ATHENS UNIVERSITY
OF ECONOMICS AND BUSINESS
DEPARTMENT OF STATISTICS
POSTGRADUATE PROGRAM

ASPECTS OF BAYESIAN MODEL AND VARIABLE SELECTION USING MCMC

By
Ioannis Ntzoufras

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Statistics

Athens, Greece
1999

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΘΕΜΑΤΑ ΜΠΕΫΖΙΑΝΗΣ ΕΠΙΛΟΓΗΣ ΥΠΟΔΕΙΓΜΑΤΩΝ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ MCMC

Ιωάννης Ντζούφρας

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διδακτορικού Διπλώματος στη Στατιστική

Αθήνα

Μαρτιος 1999

Με αγάπη στην Ιωάννα και στους γονείς μου.

Dedicated to Ioanna and my beloved parents.

Acknowledgements

I am most grateful to my supervisor, Professor Petros Dellaportas, for his academic guidance, moral support and financial assistance throughout the last three years. I would also like to thank him for encouraging me to use TEX and S+, which proved invaluable tools for my research. I should not forget to thank Dr. Jon Forster who introduced me in the world of Bayesian theory and Markov chain Monte Carlo (MCMC) methods and helped me to continue my studies in statistics.

I gratefully acknowledge all faculty members in the Department of Statistics at Athens University of Economics and Business and especially Professor Eudokia Xekalaki and Professor Ioannis Panaretos for their efforts to improve conditions of the postgraduate studies in the Department and Professor Nikos Frangos for both introducing me to the problem and providing the data for the case study of Section 1.4. I would also like to thank fellow staff and students for their help, support and creative discussions inside and outside our computer lab and particularly Stefanos Giakoumatos, Dimitris Karlis, Athina Karvounaraki, Michalis Linardakis and Ioannis Vrontos. Special thanks to Thelxi Vogiatzi for her invaluable assistance in the computer lab.

Finally, I am indebted to my parents, my brother and Ioanna Markaki for encouraging me to continue my studies, helping me to overcome difficulties when things seemed fuzzy and being patient when I was nervous due to research intractabilities. Without them this thesis would not have been completed.

Synopsis

Model selection is one of the most important problems in science. The development of a scientific theory takes the form of a mathematical model. Model selection refers to the procedure that selects hypothesized models which describe best the phenomenon under study. Traditional statistical theory involves the selection of a single model using various procedures including stepwise sequential application of significance tests or maximimization of a specified criterion which is optimal under certain conditions. On the other hand, Bayesian model selection involves the calculation of posterior model probabilities (weights). The full Bayesian model selection methods has, in the past, been restricted by difficulties in the computation of the integrals required. The recent advances in Markov chain Monte Carlo methods (MCMC) have extended the possibility to apply Bayesian model selection techniques to high dimensional problems where a large number of models may be considered. For example, in an eight way contingency table the number of hierarchical models is approximately equal to 5.6×10^{22} (Dellaportas and Forster, 1999).

The work presented in this thesis considers MCMC methods for model determination, with a general emphasis given in the popular generalised linear model formulation. Recent advances are critically reviewed and some new easy-to-use and flexible samplers are presented. Associations and connections between all existing MCMC algorithms for model selection are investigated. Moreover, a general framework for variable selection in generalised linear models is presented in detail with some associated examples. Methods for selection of link function or other structural properties are also developed and guidance of how to implement a new method, called Gibbs variable selection, using BUGS is presented. The connection of posterior odds with the information criteria, the robustness of posterior weights under different prior setups, and the presence of collinearity are investigated and discussed in detail. Finally, some calibrating methods, which eliminate the effect of prior variance on the posterior odds are proposed.

Σύνοψη

Η ανάπτυξη μιας επιστημονικής θεωρίας παίρνει συνήθως τη μορφή ενός μαθηματικού υποδείγματος. Ένα από τα σημαντικότερα προβλήματα της επιστήμης είναι η διαδικασία επιλογής ενός υποθετικού υποδείγματος που περιγράφει όσο το δυνατόν καλύτερα το υπό μελέτη φαινόμενο. Η στατιστική επιστήμη σχετίζεται παραδοσιακά με την επιλογή ενός υποδείγματος χρησιμοποιώντας μια ποικιλία από διαδικασίες στις οποίες περιλαμβάνεται η κλιμακωτή, διαδοχική εφαρμογή στατιστικών δοκιμασιών σημαντικότητας και η μεγιστοποίηση ενός προεπιλεγμένου κριτηρίου το οποίο είναι άριστο υπό ορισμένες συνθήκες. Αντίθετα, η Μπεϋζιανή επιλογή υποδειγμάτων εμπλέκει τον υπολογισμό των εκ των υστέρων πιθανοτήτων. Η πλήρης Μπεϋζιανή επιλογή υποδειγμάτων περιορίσθηκε στο παρελθόν εξαιτίας των προβλημάτων που σχετίζονται με τον υπολογισμό των ολοκληρωμάτων που απαιτούνται για την υλοποίηση του θεωρήματος του Μπεϋες. Η πρόσφατη πρόοδος των μεθόδων προσομοίωσης Monte Carlo με τη χρήση Μαρκοβιανών αλυσίδων (MCMC) για υποδείγματα πολλαπλής διάστασης επέκτειναν τη δυνατότητα εφαρμογής της Μπεϋζιανής επιλογής υποδειγμάτων σε πολύπλοκα και πολυδιάστατα προβλήματα όπου ένας μεγάλος αριθμός μοντέλων είναι υπό διερεύνηση. Για παράδειγμα σε ένα οχταπλής διάστασης πίνακα συνάφειας ο αριθμός των ιεραρχικών λογαριθμογραμμικών υποδειγμάτων είναι περίπου ίσος με 5.6×10^{22} (Dellaportas and Forster, 1999).

Η ερευνητική εργασία που εκτίθεται σε αυτή τη διατριβή μελετά της μεθόδους Monte Carlo με τη χρήση Μαρκοβιανών αλυσίδων για προσδιορισμό υποδειγμάτων με έμφαση στα δημοφιλή γενικευμένα γραμμικά μοντέλα. Παρουσιάζεται με λεπτομέρεια μια κριτική ανασκόπηση των πρόσφατων εξελίξεων και επιπλέον περιγράφεται η ανάπτυξη και η εφαρμογή ευκολόχρηστων δειγματοληπτών για τον υπολογισμό των εκ των υστέρων πιθανοτήτων των υποδειγμάτων υπό διερεύνηση. Οι σχέσεις μεταξύ όλων των υφισταμένων αλγορίθμων Monte Carlo με τη χρήση Μαρκοβιανών αλυσίδων» για επιλογή υποδειγμάτων διερευνούνται εκτενώς. Επιπλέον, παρουσιάζουμε ένα γενικό πλαίσιο για επιλογή μεταβλητών στα γενικευμένα γραμμικά υποδείγματα συνοδευόμενο με τα σχετικά παραδείγματα. Επίσης αναπτύσσουμε μεθόδους για τη επιλογή συνδετικής συνάρτησης (link function) ή άλλων δομικών χαρακτηριστικών των γενικευμένων γραμμικών μοντέλων και υποδεικνύουμε τρόπους εφαρμογής μιας νέας μεθόδου που ονομάζουμε «επιλογή μεταβλητών με τον δειγματολήπτη Gibbs» (Gibbs variable selection) με τη χρήση του λογισμικού BUGS. Η σύνδεση του εκ των υστέρων λόγου πιθανοτήτων με τα διάφορα κριτήρια επιλογής υποδειγμάτων, η ευαισθησία του εκ

των υστέρων λόγου πιθανοτήτων υπό τη χρήση διαφορετικών εκ των προτέρων κατανομών και η ύπαρξη συγγραμμικότητας εξετάζονται και συζητούνται διεξοδικά. Τέλος, παρουσιάζονται μερικοί μέθοδοι που εξαλείφουν την επίδραση της εκ των προτέρων διακύμανσης πάνω στον εκ των υστέρων λόγο πιθανοτήτων ούτως ώστε να χρησιμοποιήσουμε εκ των προτέρων κατανομές με όσο μεγάλη διακύμανση επιθυμούμε χωρίς να αλλάζουν σημαντικά οι πιθανότητες επιλογής του κάθε υποδείγματος.

Contents

List of Tables

List of Figures

Abbreviations

ABBREVIATION	**DESCRIPTION**
AIC	Akaike information criterion
BIC	Bayes information criterion
CCVLS	Carlin and Chib variable and link selection sampler
CCVS	Carlin and Chib variable selection Gibbs sampler
GLM	Generalised linear model
GVS	Gibbs Variable Selection
KM	Kuo and Mallick (1998) Gibbs sampler
LR	Likelihood Ratio
MCC	Metropolised Carlin and Chib sampler
MC^3	Markov chain Monte Carlo model composition
MCMC	Markov chain Monte Carlo
PR	Prior density ratio
PSR	Pseudoprior ratio
RJ	Reversible jump Metropolis sampler
RSS	Residual sum of squares
SS	Posterior residual sum of squares
SSVS	Stochastic Search Variable Selection
SSVLS	Stochastic Search Variable and Link Selection

Notation

NOTATION	**DESCRIPTION**
$\mathbf{0}_d$	$d \times 1$ vector of zeros
$\mathbf{1}_d$	$d \times 1$ vector of ones
$\mathcal{A}$	Set of models selected by first condition in Occam's Window
$\mathcal{A}'$	Set of models selected by Occam's Window
$\boldsymbol{A}_j^*$	Quantity used in Clyde sampler (page 79)
A_m	Quantity used in Carlin and Chib sampler (page 71)
$a_\imath$	Parameter for $\imath$ accident year in Section 1.4
a_τ	Prior parameter for σ^{-2}
α	Acceptance probability in Metropolis algorithms
α_1^*, α_2^*	Quantities used in Chapter 6
$a_i(\phi)$	Function of ϕ in exponential family for i observation
$\mathcal{B}'$	Set of models selected by second condition in Occam's Window
$\mathcal{B}_m$	Set of possible values of $\boldsymbol{\beta}_{(m)}$
B_{01}	Bayes factor of model m_0 versus m_1
$B(a,b)$	Beta distribution with parameters a, b
b_0	Constant model parameter used in case study of Section 1.4
$b_\jmath$	Parameter for $\jmath$ delay year in Section 1.4
$b_\jmath^*$	Parameter for $\jmath$ delay year for multinomial stage in Section 1.4
b_τ	Prior parameter for σ^{-2}
$b(\vartheta)$	Function of ϑ in exponential family
$\boldsymbol{\beta}$	Model parameters of full model
$\boldsymbol{\beta}_{(L)}$	Model parameters of full model for link L
$\boldsymbol{\beta}^*$	All parameter vectors $\boldsymbol{\beta}_{(L)}$ for all $L \in \boldsymbol{\mathcal{L}}$
$\boldsymbol{\beta}_j$	Sub-vector of model parameters for term j
$\boldsymbol{\beta}_{j,(L)}$	Sub-vector of model parameters for term j and link L
$\boldsymbol{\beta}_{\setminus j}$	Vector of all model parameters excluding term j
$\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$	Parameter vector of model $\boldsymbol{\gamma}$

$\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}$	Parameter vector of model $\boldsymbol{\gamma}$ and L
$\boldsymbol{\beta}_{(m)}$	Parameter vector of model m
$\boldsymbol{\beta}_{(m,x_j)}$	Parameter vector of model with response variable X_j and design matrix $\boldsymbol{X}_{(m)}$
$\hat{\boldsymbol{\beta}}_{(m)}$	Maximum Likelihood Estimates of parameter vector of model m
$\tilde{\boldsymbol{\beta}}_{(m)}$	Posterior mean of parameter vector of model m
$\breve{\boldsymbol{\beta}}_{(m)}$	Posterior mode of parameter vector of model m
β_0	Constant model parameter
$\mathcal{C}$	Set of models used in Up and Down algorithms of Occam's Window
c^2	Variance prior parameter controlling the flatness
$c(y, a_i(\phi))$	Function used in exponential family
$\boldsymbol{\gamma}$	Variable/term indicator vector
$\mathcal{D}$	Set of distributions under consideration
D	Distributional indicator
$D_{(m)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)$	Block diagonal matrix of dimension $d(m) \times d(m)$ and diagonal elements the matrices $\boldsymbol{X}_j^T\boldsymbol{X}_j$
$\boldsymbol{D}_{\boldsymbol{\gamma}}$	SSVS prior parameter
d	Dimension of full model
d_j	Dimension of j term
$d(m)$	Dimension of model vector $\boldsymbol{\beta}_{(m)}$ or model m
Δ	Prediction parameter
δ	Negative binomial indicator
δ_j	SSVS value of practical significance for j term
$\delta^*(\mu_0)$, $\delta^{**}(\mu_0)$	Linear coefficients used in Taylor approximation (see page 139)
$E_\imath$	Exposure corresponding to $\imath$ accident year (Section 1.4)
$\mathcal{E}_i$	Indicator parameter used in probit model (see page 89)
ϵ	Varying parameter used in negative binomial formulation
$\varepsilon_\imath$	State space variance for $\imath$ accident year in Section 1.4
F	Penalty imposed to maximised log-likelihood function for each additional parameter used
$FBF_{b,01}$	Fractional Bayes factor of model m_0 versus m_1 with fractional parameter b

F_j	$\boldsymbol{y}^T\boldsymbol{X}_j\left(\boldsymbol{X}_j^T\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j^T\boldsymbol{y} = \hat{\boldsymbol{\beta}}_j^T\boldsymbol{X}_j^T\boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j$
$F_{\chi^2_d}(*)$	Probability function of χ^2_d distribution
$G(a,b)$	Gamma distribution with mean a/b and variance a/b^2
$\mathcal{G}_j$	Region of insignificance in SSVS priors for multidimensional terms
$g(*)$	Link function
$g_L(*)$	Link function that corresponds to L indicator
$g_\lambda(*)$	Link function that corresponds to λ parameter
$\Gamma(a)$	Gamma function with parameter a
$\boldsymbol{H}$	Diagonal matrix used in Fisher information matrix of the full model
$\hat{\boldsymbol{H}}$	Diagonal matrix used in observed information matrix of the full model
$\boldsymbol{H}_{(m)}$	Diagonal matrix used in Fisher information matrix of model m
$\hat{\boldsymbol{H}}_{(m)}$	Diagonal matrix used in observed information matrix of model m
h_i	Element of matrix $\boldsymbol{H}$
$h_{L,L'}(*)$	Transforming function from link L to L' used in reversible jump
$h_{m,m'}(*)$	Transforming function from model m to m' used in reversible jump
$\zeta_\imath$	State space variance for $\imath$ accident year in Section 1.4
$\boldsymbol{\eta}$	Vector of linear predictors
$\boldsymbol{\eta}_j^*$	Residual linear predictor when all variables except j are known
$\boldsymbol{\eta}^L$	Vector of linear predictors for L link function
η_i	Linear predictor for i subject/observation
η_i^L	Linear predictor for i subject/observation and L link
$\{\boldsymbol{\eta}_j^*\}_i$	i element of $\boldsymbol{\eta}_j^*$
$\mathcal{I}_{\boldsymbol{\beta}_{(m)}}$	Minus the inverse of Fisher information matrix for model m
$\mathcal{I}_{\hat{\boldsymbol{\beta}}_{(m)}}$	Minus the inverse of observed Fisher information matrix for model m
IC_{01}	Information criterion of model m_0 versus m_1
$\boldsymbol{I}_d$	Identity matrix of dimension $d \times d$
$I(A)$	Indicator function taking value one if A is true and zero otherwise
IBF_{01}	Intrinsic Bayes factor of model m_0 versus m_1
$IG(a,b)$	inverse gamma distribution with mean $b/(a-1)$
i	Observation indicator

$\imath$	Indicator for accident year in Section 1.4
$inf_{\imath\jmath}$	Inflation corresponding to $\imath$ accident and $\jmath$ delay year (Section 1.4)
$\boldsymbol{\theta}_m$	Set of parameters for model m
θ	Dispersion parameter used in negative binomial
ϑ	Exponential family natural canonical parameter
$\boldsymbol{J}_d$	Matrix of dimension $d \times d$ with $[\boldsymbol{J}]_{kl} = 1$ for all k, l
j	Term/variable indicator
$\jmath$	Indicator for delay year in Section 1.4
$j(m, m')$	Proposal probability of jumping from model m to model m'
K^2	Variance inflation parameter used in outlier identification
$K(f, g)$	Kullback-Leibler distance between f and g (see 38)
$K_m^{m_0}$	Kullback-Leibler based predictive criterion of models m and m_0 (page 47)
k^2, k_j^2	Small variance divisor used in SSVS
k^*	Constant parameter used in Occam's window
κ	Prior parameter used in Chapter 6
$\mathcal{L}$	Set of link functions under consideration
L, L', L'', ℓ	Link indicator variables
L_m^2	Ibrahim and Laud (1994) predictive criterion (see page 46)
LS	Logarithmic scoring rule for Bayesian model averaging
LS_m	Logarithmic scoring rule for model m
$l(\boldsymbol{\beta}_{(m)})$	Likelihood function of model m
$l(\boldsymbol{y}, \boldsymbol{y}^*, w_1, w_2 \| m)$	Weighted likelihood of model m giving weight w_1 to each data point of $\boldsymbol{y}$ and weight w_2 to each data point of $\boldsymbol{y}^*$
λ_i	Mean value of i subject in Poisson models
$\lambda_{\imath\jmath}$	Expected value of claim counts for $\imath$ accident and $\jmath$ delay year used in Section 1.4
$\mathcal{M}$	Set of all models under consideration
$\|\mathcal{M}\|$	Number of all models under consideration
M_m^*	Laud and Ibrahim (1995) predictive criterion (see page 46)
m, m', m_k, m_l	Model indicators
$\boldsymbol{\mu}$	$n \times 1$ vector of model expected values

$\boldsymbol{\mu}_{\beta_j}$	Prior mean of $\boldsymbol{\beta}_j$
$\bar{\boldsymbol{\mu}}_{\beta_j}$	Proposal mean of $\boldsymbol{\beta}_j$
$\boldsymbol{\mu}_{\beta_{(\gamma,L)}}$	Prior mean of $\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}$
$\bar{\boldsymbol{\mu}}_{\beta_{(\gamma,L)}}$	Proposal mean of $\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}$
$\boldsymbol{\mu}_{\beta_{(m)}}$	Prior mean of $\boldsymbol{\beta}_{(m)}$
$\mu_{\imath\jmath}$	Expected value of inflation adjusted claim amount for $\imath$ accident and $\jmath$ delay year (see Section 1.4)
$\bar{\boldsymbol{\mu}}_{\beta_{(m)}}$	Proposal mean of $\boldsymbol{\beta}_{(m)}$
μ_i	Model expected value for i observation
$N(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution of d dimension with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
N_i	Total number of trials in binomial distribution
n	Sample Size or dimension of $\boldsymbol{y}$ vector
n_0	Sample Size of prior data / dimension of $\boldsymbol{y}^*$ vector
$n_{\imath\jmath}$	Claim counts for $\imath$ accident and $\jmath$ delay year (Section 1.4)
$nb(m)$	Set of neighbour model of m
$\|nb(m)\|$	Number of neighbour models of m
ν	Term indicator
$\Xi(*)$	Quantity used in Metropolised Gibbs link selection (144)
ξ_1, ξ_2	Prior parameters used alternative prior setting
O_j	Conditional posterior odds in Gibbs variable selection steps
O'_j	Quantity used in Clyde (1999) method (see page 79)
o_i	Mean value for outliers
PBF_{01}	Posterior Bayes factor of model m_0 versus m_1
$\boldsymbol{P}_{\boldsymbol{\gamma}}$	$\boldsymbol{X}_{(\boldsymbol{\gamma})}\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1}\boldsymbol{X}^T_{(\boldsymbol{\gamma})}$
$\boldsymbol{P}_m$	$\boldsymbol{X}_{(m)}\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\right)^{-1}\boldsymbol{X}^T_{(m)}$
$p, \|\mathcal{V}\|$	Number of all terms/variables under consideration
p_i	Binomial success probability for i subject/observation

p'_j	Multinomial probability to have delay $j-1$ years (Section 1.4)
p''_j	Binomial probability used in Gibbs sampling of Section 1.4
π	Common probability of one term to be included in the model ($f(\gamma_i)=\pi$)
π_j	Probability of j term to be included in the model ($f(\gamma_i)=\pi_j$)
PO_{01}	Posterior odds of model m_0 versus model m_1
$POLS^q_{01}$	Posterior odds of model m_0 versus model m_1 at the limit of significance q
PrO	Prior odds for each term included in model ($PrO=\pi/(1-\pi)$)
$\mathcal{R}$	Set of transformations of Y under consideration
$\boldsymbol{R}_{\boldsymbol{\gamma}}$	SSVS prior variance parameter
RSS_m	Residual sum of squares of model m
r	Dimension of data table in case study of Section 1.4
$\boldsymbol{Q}_{\boldsymbol{v}}$	Diagonal precision matrix depending on outliers
q	Significance level in significance tests
$q(\boldsymbol{u}\|\boldsymbol{\beta}_{(m)},m,m')$	Proposal density of additional parameters given that we are in $(\boldsymbol{\beta}_{(m)},m)$ and propose move to model m'
ρ	Parameter used in Box-Cox transformation
ϱ	Parameter used continous link approaches
$\mathcal{S}$	Set of structural properties of models
$\bar{\boldsymbol{S}}_j$	Proposal Covariance for j term
$\bar{\boldsymbol{S}}_{j,(L)}$	Proposal Covariance for j term and L link
$\bar{\boldsymbol{S}}_{(m)}$	Proposal Covariance for model m
$SS_m, SS_{\boldsymbol{\gamma}}$	Posterior sum of squares of model m or $\boldsymbol{\gamma}$
$SS_{\boldsymbol{T},\boldsymbol{\gamma}}$	Posterior sum of squares of model $\boldsymbol{\gamma}$ when using heteroscedastic structure
$SS_{\boldsymbol{v},\boldsymbol{\gamma}}$	Posterior sum of squares of model $\boldsymbol{\gamma}$ in outlier identification
$SS_{\boldsymbol{Z},\boldsymbol{\gamma}}$	Posterior sum of squares of model $\boldsymbol{\gamma}$ in Probit regression
$SS^*_{\boldsymbol{\gamma}}$	SSVS posterior sum of squares of model $\boldsymbol{\gamma}$ as defined in page 80
s	Model structural indicator
s_j	Sample variance of X_j
s_y	Sample variance of Y
$\boldsymbol{\Sigma}_j$	Prior covariance matrix for term j
$\tilde{\boldsymbol{\Sigma}}_j$	Posterior covariance matrix of $\boldsymbol{\beta}_j$

Σ_j	Prior variance one-dimensional term j
$\boldsymbol{\Sigma}_{j,(L)}$	Prior covariance matrix for term j and L link
$\boldsymbol{\Sigma}_{(m)}$	Prior covariance matrix for model m
$\tilde{\boldsymbol{\Sigma}}_{(m)}$	Posterior covariance matrix of $\boldsymbol{\beta}_{(m)}$
$\sigma^2_{b_0}$, $\sigma^2_{a_i}$, $\sigma^2_{b_j}$	Prior variances for the corresponding parameters in Section 1.4
$\boldsymbol{T}$	Heteroscdastic covariance structure in linear models
$TE_L(\mu_0)$	Taylor expansion linear approximation of link function $g_L(*)$ round μ_0
T_i	Total number of counts for i accident year (Section 1.4)
t	Denoting time or iteration in MCMC
$U(a,b)$	Uniform distribution with range from a to b
$\boldsymbol{u}$	Vector of proposed parameters in reversible jump
$u'(m,m')$	Utility function corresponding to selection of model m' when m is correct
$u'_{m,m'}$	Constant utility corresponding to selection of model m' when m is correct
$\Phi(*)$	Standardised normal distribution function
ϕ	Dispersion parameter of exponential family
$\varphi(*)$	Standardised normal density function
X^s	Standardised X
$\boldsymbol{x}_i^T$	Data vector for observation i
$\boldsymbol{X}$	Data matrix of full model
X_j	Explanatory variables, factors or interaction terms
$\boldsymbol{X}_j$	Data matrix for term/variable j
$\bar{x}_j$	Sample mean of X_j
$\boldsymbol{X}_{(\boldsymbol{\gamma})}$	Data matrix for model $\boldsymbol{\gamma}$
$\boldsymbol{X}_{(m)}$	Data matrix for model m
$\boldsymbol{X}^*_{(m)}$	Prior data matrix for model m
$\boldsymbol{X}_{(\boldsymbol{v},\boldsymbol{\gamma})}$	Data matrix for model $\boldsymbol{\gamma}$ excluding outliers according to $\boldsymbol{v}$
ψ	Penalty function imposed to the maximised likelihood function
Y	Response variable
Y^s	Standardised Y
Y_{ij}	Claim amounts for i accident and j delay year (Section 1.4)
$\boldsymbol{y}$	$n \times 1$ response data vector

$\boldsymbol{y}^*$	$n_0 \times 1$ response of prior data vector
$\boldsymbol{y}(l)$	Training sample used in intrinsic Bayes factor
$\boldsymbol{y}(\backslash l)$	Data used for model selection in intrinsic Bayes factor
$\boldsymbol{y}(\rho)$	Box Cox Transform data of Y with parameter ρ
$\boldsymbol{y_v}$	Data of Y that are not outliers
$\boldsymbol{y}_{\backslash \boldsymbol{v}}$	Data of Y that are outliers
$\boldsymbol{y}$	$n \times 1$ response data vector
$\boldsymbol{y}$	$n \times 1$ response data vector
$\bar{y}$	Sample mean of Y
$\Upsilon_{\imath\jmath}$	Inflation adjusted claim amounts for $\imath$ accident and $\jmath$ delay year
$\boldsymbol{Z}$	Vector of values for latent variables used in probit regression
$\boldsymbol{z}$	Replicated data used in predictive criteria
z_q	q quantile of standardised normal distribution
$\mathcal{V}$	Set of all terms/variables under consideration
$\mathcal{V}(j)$	Set of all variables included in term j
$\mathcal{V}(m)$	Set of all terms/variables included in model m
$\boldsymbol{V}_{(m)}$	Prior matrix used in covaariance ($\boldsymbol{\Sigma}_{(m)} = c^2 \boldsymbol{V}_{(m)}$)
v	Negative binomial indicator variable
$\boldsymbol{v}$	Vector of v_i outlier indicator variables
v_i	Outlier indicator variable of i observation
v_m	Binary indicator variable for model m
w_i	Exponential family weights ($a_i(\phi) = \phi / w_i$)
w_i^*	Weights used in importance sampling
$\Omega_{\imath\jmath}$	Parameter used in Gibbs sampling of Section 1.4

Introduction

One of the most important issues in statistical science is the construction of probabilistic models that represent, or sufficiently approximate, the true generating mechanism of a phenomenon. The arbitrary construction of such models may possibly include useless information for the description of the phenomenon under study. Model selection is the procedure that decides which probabilistic structure we should finally select from a specified set of models. All model selection procedures try to balance two different notions: goodness of fit and parsimony. The first notion refers to the procedure of selecting models that describe the available data as good as possible while the second notion refers to the procedure of avoiding unecessary complication of the model.

Although a model selection procedure seems to be defined clearly, the identification of a mathematical procedure for the selection of 'good' models is still problematic even in the simple case of covariate selection, that is, selection of variables that influence a response variable Y under study. Particularly in variable selection the most broadly used methods are the stepwize procedures which consist of a sequential application of single significance tests. The simplest arguement against stepwize methods is that their distribution cannot be identified and is by no means the same to the distribution of the single significance test used (see, for example, Miller, 1984). Moreover, significance tests cannot discriminate between non-nested models and therefore between models with different distributional structures.

A variety of alternative model selection criteria have been presented in statistical literature. These criteria select the model which maximizes a quantity usually expressed as the maximum log-likelihood minus a penalty function which depends on the dimensionality of the model. The most popular criteria were introcuded by Akaike (AIC, 1973), Mallows (C_p, 1973) and Schwarz (BIC, 1978). The large variety of different penalty functions (which are optimal under certain conditions) have made practitioners to wonder which of these criteria should consider under different circumstances. Even if we limit ourselves to the popular AIC or BIC, the model selection proccess will usually result to different models (BIC selects more parsimonious models). A further problem that we face when using any model selection criterion is the large number of models that we need to consider. For example if we have a normal linear regression model with 15 covariates then the models under consideration are 2^{15}. Calculating the selected criterion over all these models is highly inefficient. Alternative procedures, such as stepwize type methods, may not trace the model which maximizes the criterion used since, in collinear cases, some 'good' models will not be visited at all.

On the other hand, Bayesian model selection offers solutions to the above problems with the use of modern computing tools such as Markov chain Monte Carlo (MCMC) techniques which explore the model space, trace 'good' working models and estimate their prosterior probability (or more appropriate their posterior weight) based on both assigned prior beliefs and observed data. The main problem which has prevented Bayesian model selection techniques from being a broadly accepted solution for model selection problems is the sensitivity of posterior weights to different prior distributions. No prior distribution can be thought as entirely non-informative since proper approximately flat priors (thought to be approximately non-informative for a given model) fully support the simplest, in terms of dimensionality, model.

This thesis attempts to enlighten some aspects of Bayesian model selection including MCMC methods and problems caused by prior misspecification. The thesis is organized into eight chapters. The first provides introductory details for Bayesian model inference through MCMC algorithms and a case study in an actuarial problem. The second chapter gives a general overview and discussion of model selection techniques including stepwize methods, information based criteria and Bayesian techniques. A critical review of available prior distributions and MCMC techniques used in Bayesian model selection are presented in the third chapter. Some easy-to-use samplers for variable selection are developed in Chapter 4. Connections between widely used samplers for model selection are also presented, as well as a general implementation framework for covariate selection in generalised linear models. Advanced samplers for the selection of structural properties such as response distribution, link function and residual identification are developed in Chapter 5 while further topics of Bayesian model selection, such as the effect of different prior distributions on the posterior weights or the problems caused by collinear covariates, are examined in Chapter 6. Associations between Bayesian model selection methods and information criteria are also presented together with some proposed methods for specification of prior distributions via determination of a prefixed penalty function. The thesis is concluded with a detailed guide and illustrated examples for the implementation of the variable selection sampler introduced in this thesis and a discussion including the directions for further research in Chapters 7 and 8.

Chapter 1

Model Based Bayesian Inference via Markov Chain Monte Carlo

1.1 Definition of Statistical Models

Assume that a random variable Y, usually called response, follows a probabilistic rule $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the parameter vector. Consider an i.i.d. sample $\boldsymbol{y}^T = [y_1, \ldots, y_n]$ of size n of this variable. The joint distribution

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$$

is called the likelihood of the model and contains all the available information provided by the sample.

Usually models are constructed in order to assess or interpret causal realationships of the response variable Y with other characteristics expressed in variables X_j, $j \in \mathcal{V}$, usually called covariates; j indicates a covariate or model term and $\mathcal{V}$ the set of all terms under consideration. In such cases these explanatory variables are linked with the response variables via a deterministic function and part of the parameter vector is substituted by alternative parameters (noted by $\boldsymbol{\beta}$) that usually encapsulate the effect of each covariate on the response variable. For example in a regression model with $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ the parameter vector is given by $\boldsymbol{\theta}^T = [\boldsymbol{\beta}^T, \sigma^2]$.

1.2 Model Based Bayesian Inference

Bayesian theory differs from classical statistical theory since it considers any unknown parameter as random variable and, for this reason, each of these parameters should have a prior distribution. Therefore, interest lies in calculating the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})$ of the unknown parameters which incorporate both prior $[f(\boldsymbol{\theta})]$ and data $[f(\boldsymbol{y}|\boldsymbol{\theta})]$ information. When covariates are incorporated in the model formulation, interest usually lies in calculating the posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{y})$ rather than $f(\boldsymbol{\theta}|\boldsymbol{y})$.

The moments of the posterior distribution may be used as summary descriptive measures for inference. If no prior information is available a wide range of 'non-informative' vague priors may be used; for details see Kass and Wasserman (1996) and Yang and Berger (1996). In many cases the posterior distribution is intractable. In the past, intractability was avoided via 'conjugate' prior distributions. These distributions have the nice property of resulting to posteriors of the same distribution family. Extensive illustration of conjugate priors is provided by Bernando and Smith (1994). In cases that conjugate priors are considered to be unrealistic or are unavailable, asymptotic approximations such as Laplace approximation may be used (see, for example, Tierney and Kadane, 1986, Tierney *et al.* , 1989 and Erkanli, 1994) or numerical integration techniques (see, for example, Evans and Swartz, 1996). In recent years, massive development of computing facilities has made MCMC techniques popular. These techniques generate samples from the posterior distribution. MCMC enabled Bayesians to use highly complicated models and estimate the posterior densities with accuracy. These methodologies are briefly described in the next section.

1.3 Markov Chain Monte Carlo Methods

MCMC methodology is a very powerful computational tool which has recently become popular in the statistical community. The main reason for its popularity is its ability to evaluate (indirectly) high dimensional integrals involved in the Bayesian implementation of statistical models describing common real life problems. MCMC methods were initially introduced by Metropolis *et al.* (1953), but were made popular in statistical science after the publications of Gelfand and Smith (1990) and Gelfand *et al.* (1990). Extensive details of the use of MCMC methods are given by Gilks *et al.* (1996). Finally, the BUGS software developed by Spiegelhalter *et al.* (1996a,b,c) provides an easy-to-use program for applying MCMC methods in Bayesian modelling.

A Markov chain is a stochastic process $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(t)}\}$ such that $f(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}, \cdots, \boldsymbol{\theta}^{(1)}) = f(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})$. That is, the distribution of $\boldsymbol{\theta}$ in time $t+1$ given all the preceding $\boldsymbol{\theta}$ (for times $t,\ t-1,\ \ldots,\ 1$) depends only on $\boldsymbol{\theta}^{(t)}$. Moreover, $f(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})$ is independent of time t. Finally, when the Markov chain is irreducible, aperiodic and positive recurrent, as $t \to \infty$ the distribution of $\boldsymbol{\theta}^{(t)}$ tends to its equilibrium distribution which is independent of the initial $\boldsymbol{\theta}^{(0)}$; for details see Gilks *et al.* (1996).

In order to generate a sample from $f(\boldsymbol{\theta}|\boldsymbol{y})$ we must construct a Markov chain with two desired properties. First, $f(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})$ should be 'easy to generate from' and, second, the equilibrium distribution of the selected Markov chain should be our target posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$.

We construct a Markov chain with the above requirements, then we select an initial value $\boldsymbol{\theta}^{(0)}$ and generate values until the equilibrium distribution is reached. The next step is to cut off the first t_0 observations and take as a sample $\{\boldsymbol{\theta}^{(t_0+1)}, \boldsymbol{\theta}^{(t_0+2)}, \cdots, \boldsymbol{\theta}^{(t_0+t)}\}$. Convergence of the MCMC can be checked by various methods; for details see Cowles and Carlin (1996) and Brooks and Roberts (1997). CODA software, which applies certain tests to check MCMC convergence, is provided by Best *et al.* (1995).

Two are the most popular MCMC methods: Metropolis-Hastings algorithm (Metropolis *et al.* , 1953, Hastings, 1970) and the Gibbs sampler (Geman and Geman, 1984).

1.3.1 The Metropolis-Hastings Algorithm

Metropolis *et al.* (1953) introduced MCMC methods with *Metropolis* algorithm. Seventeen years later, Hastings (1970) generalised the original method in what is known as *Metropolis-Hastings* algorithm. The latter is considered as the general form of any MCMC method. Green (1995) further generalised Metropolis-Hastings algorithm by introducing *reversible jump Metropolis-Hastings* algorithms for sampling from parameter spaces with different dimension.

In Metropolis-Hastings algorithm we follow iteratively three steps:

1. Generate $\boldsymbol{\theta}'$ from a proposal distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

2. Calculate
$$\alpha = min\left(1, \frac{f(\boldsymbol{\theta}'|\boldsymbol{y})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}')}{f(\boldsymbol{\theta}^{(t)}|\boldsymbol{y})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)})}\right).$$
[Both $f(.)$ and $q(.)$ do not require their normalising constants because they cancel out].

3. Update $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}'$ with probability α, otherwise set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

Special cases of Metropolis-Hastings are the *Metropolis* algorithm, *random walk Metropolis*, *Independence Sampler*, *single component Metropolis-Hastings* and the *Gibbs sampler*. In Metropolis algorithm, Metropolis *et al.* (1953), only symmetric proposals were considered, that is $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)}) = q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}')$. Random walk Metropolis is a special case with $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)}) = q(|\boldsymbol{\theta}' - \boldsymbol{\theta}^{(t)}|)$. Both cases result in
$$\alpha = min\left(1, \frac{f(\boldsymbol{\theta}'|\boldsymbol{y})}{f(\boldsymbol{\theta}^{(t)}|\boldsymbol{y})}\right).$$
A usual proposal of this type is $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)}) \equiv N(\boldsymbol{\theta}^{(t)}, \mathbf{S})$. The covariance matrix $\mathbf{S}$ controls the convergence speed of the algorithm.

Independence sampler is a Metropolis Hastings algorithm where the proposal distribution does not depend on the current state $\boldsymbol{\theta}^{(t)}$ of the chain. This sampler can be used when a good approximation of the posterior distribution is known. In many cases a good independent proposal may be given by Laplace approximation (Tierney and Kadane, 1986, Tierney *et al.* , 1989, Erkanli, 1994).

Finally, in single component Metropolis-Hastings only one component at each time is updated. In each step, a candidate value of the jth component of the vector $\boldsymbol{\theta}$, θ'_j, is proposed by $q_j(\theta'_j|\boldsymbol{\theta}^{(t)})$. Gibbs sampler is a special case of this algorithm and will be discussed in detail in the following section. Other variations of Metropolis Hastings have been developed; for a detailed description see Chib and Greenberg (1995) and Gilks *et al.* (1996).

1.3.2 Gibbs Sampler

Gibbs sampler was introduced by Geman and Geman (1984). In this algorithm we update one component in each step from the corresponding conditional posterior. Given a particular state of the chain $\boldsymbol{\theta}^{(t)}$ we we have the following steps

$$
\begin{array}{lll}
\theta_1^{(t+1)} & \text{from} & f(\theta_1|\theta_2^{(t)},\theta_3^{(t)},\cdots,\theta_p^{(t)},\boldsymbol{y}),\\
\theta_2^{(t+1)} & \text{from} & f(\theta_2|\theta_1^{(t+1)},\theta_3^{(t)},\cdots,\theta_p^{(t)},\boldsymbol{y}),\\
\theta_3^{(t+1)} & \text{from} & f(\theta_3|\theta_1^{(t+1)},\theta_2^{(t+1)},\cdots,\theta_p^{(t)},\boldsymbol{y}),\\
\vdots\ \vdots\ \vdots & & \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots\\
\theta_p^{(t+1)} & \text{from} & f(\theta_p|\theta_1^{(t+1)},\theta_2^{(t+1)},\cdots,\theta_{p-1}^{(t+1)},\boldsymbol{y}),
\end{array}
$$

where p is the number of components of the parameter vector $\boldsymbol{\theta}$. The generation from $f(\theta_j|\boldsymbol{\theta}_{\setminus j},\boldsymbol{y}) = f(\theta_j|\theta_1^{(t+1)},\cdots,\theta_{j-1}^{(t+1)},\theta_{j+1}^{(t)},\cdots,\theta_p^{(t)},\boldsymbol{y})$ is relatively easy since it is a univariate distribution and can be written as $f(\theta_j|\boldsymbol{\theta}_{\setminus j},\boldsymbol{y}) \propto f(\boldsymbol{\theta}|\boldsymbol{y})$ where all the variables except θ_j are held constant at their given values. Gibbs sampler is a special case of single component Metropolis-Hastings algorithm since, when the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)})$ equals to the full conditional posterior distribution $f(\theta_j|\boldsymbol{\theta}_{\setminus j},\boldsymbol{y})$, we have $\alpha = 1$ and therefore we always accept the proposed move. More detailed description of the Gibbs sampler is given by Casella and George (1992) and Gilks *et al.* (1996) while applications of Gibbs sampling are given by Gelfand and Smith (1990), Gelfand *et al.* (1990) and Smith and Roberts (1993).

1.4 Case Study: Bayesian Modelling of Outstanding Liabilities Incorporating Claim Count Uncertainty

Here we present a complete and detailed actuarial case study in order to illustrate how we can exploit the possibilities and solutions offered by Bayesian inference and MCMC techniques. Summary of the implementation of this case study is also available in the form of research paper; see Ntzoufras and Dellaportas (1998).

1.4.1 Background of the Case Study

Insurance companies often do not pay the outstanding claims as soon as they occur. Instead, claims are settled with a time delay which may be years or, in some extreme cases, decades. Reserving for outstanding claims is of central interest in actuarial practice and has attracted the attention of many researchers because of the challenging stochastic uncertainties involved.

In this section, we investigate possible extensions of models which deal with outstanding liabilities. We focus in cases for which the outstanding claims will be considered to be only the 'reported but not settled' claims. Thus, other important cases such as claims settled with sub-payments or 'incurred but not reported' claims (including claims that have been already settled but may be reopened) are not considered.

Mathematically, the problem can be formulated as follows. There exist data with a structure given by Table 1.1. A_i, $i = 1, 2, \ldots, r$ denote the accident years and B_j, $j = 1, 2, ..., r$ denote the years that the claim was settled. For example, the cell ij contains the amount Y_{ij} that the company paid with a delay of $j - 1$ years for accidents happened during the year i. Moreover, the counts of claims for each cell are also given for certain insurance claims. The claim counts have the same triangular form as the claim amounts; in Table 1.2, n_{ij} represents claim counts that an insurance company paid with a delay of $j - 1$ years for accidents originated at year i and T_i denotes the total number of accidents. Finally, the inflation factor for each cell inf_{ij} which is used to deinflate the claim amounts, is also assumed to be known.

		B				
		B_1	B_2	$\cdots$	B_{r-1}	B_r
	A_1	Y_{11}	Y_{12}	$\cdots$	$Y_{1,r-1}$	Y_{1r}
	A_2	Y_{21}	Y_{22}	$\cdots$	$Y_{2,r-1}$	
A	$\vdots$	$\vdots$	$\vdots$			
	A_{r-1}	$Y_{r-1,1}$	$Y_{r-1,2}$			
	A_r	Y_{r1}				

Table 1.1: Structure of Outstanding Claim Amounts Data.

Many models and techniques have been proposed for the prediction of the lower triangles of Tables 1.1 and 1.2. Two of the most broadly used models are the log-linear and the log-normal anova-type models; see Taylor and Ashe (1983), Renshaw and Verrall (1994), Renshaw (1989) and Verrall (1991, 1993, 1996). Renshaw (1994) found accurate approximations of the square root prediction error for the log-normal model. A short review of more advanced models for this problem is given by Haberman and Renshaw (1996). Recently,

		B					
		B_1	B_2	$\cdots$	B_{r-1}	B_r	
	A_1	n_{11}	n_{12}	$\cdots$	$n_{1,r-1}$	n_{1r}	T_1
	A_2	n_{21}	n_{22}	$\cdots$	$n_{2,r-1}$		T_2
A	$\vdots$	$\vdots$	$\vdots$				$\vdots$
	A_{r-1}	$n_{r-1,1}$	$n_{r-1,2}$				T_{r-1}
	A_r	n_{r1}					T_r

Table 1.2: Structure of Outstanding Claim Counts Data.

Verrall (1996) investigated generalised additive models. Dynamic models (also called state space models) were also proposed; De Jong and Zehnwirth (1983) made a general investigation of state space models in reserving claims problem and Verrall (1989, 1994) investigated a state space model which can be viewed as an extension to the log-normal models. From a Bayesian point of view, Verrall (1990) produced Bayes estimates for the log-normal model. Makov *et al.* (1996) provide a short review of Bayesian methods and description of applying MCMC in the problem of outstanding claims. Bayesian methods were also used by Haastrup and Arjas (1996) for estimating claim counts and amounts in individual claim data, Jewell (1989) and Alba *et al.* (1997) for estimating claim counts. Empirical Bayes estimates were also obtained by Verrall (1989, 1990). Further work in estimating claim amounts or claim counts is published by Norberg (1986), Hesselager (1991), Neuhaus (1992) and Lawless (1994).

In this case study we propose a Bayesian approach to investigate various models for the outstanding claims problem. The use of the Bayesian paradigm did not emanate from the need to use prior information as in Verrall (1990) but, rather, from its computation flexibility that allows us to handle complex models. MCMC sampling strategies are used to generate samples for each posterior distribution of interest. A key feature in the modelling approach we propose, is the simultaneous use of data of Tables 1.1 and 1.2. In this way we are essentially modelling 'payments per claim finalized' (PPCF) in delay year j having origin in year i; see Taylor and Ashe (1983). This has the advantage that increase in accidents, which is expressed via increase of claim counts in Table 1.2, results in the increase of total claims in Table 1.1. Modelling of both claim counts and amounts has been advocated in the past; see Norberg (1986), Hesselager (1991), Neuhaus (1992), Haastrup and Arjas (1996) for application in individual claim data and Taylor and Ashe (1983) for aggregated data.

We model the claim counts for year i using a multinomial distribution when T_i is known. If T_i is not known, this assumption is easily relaxed and the claim counts data can be modelled as Poisson distributed.

In the next section we briefly describe the Bayesian theory and the MCMC methodology. Section 1.4.3 presents the models under consideration. In Section 1.4.4 an illustrated example with data from a Greek insurance company is presented. Finally, a brief discussion is given in Section 1.4.5 while computational details are given in Section 1.4.6.

1.4.2 Bayesian Modelling via MCMC

We adopt Bayesian theory and MCMC methodology in order to develop new models in estimating future liabilities. Technicalities involved MCMC methods are given in Section 1.3. To avoid extended discussion on MCMC, the following section focuses only in the modelling aspects of the problem and implementation details for the models proposed (form of the full conditional densities, Metropolis Hastings sampling etc.) are given in Section 1.4.6. Technicalities used include sampling from the usual log-concave densities met in generalised linear models (Dellaportas and Smith, 1993), MCMC strategies for problems with constrained parameter problems (Gelfand *et al.* , 1992), ways to handle the missing values problem (Gelfand *et al.* , 1990), and finally, usual transformation of the MCMC output to the actual parameters of interest. For the latter, let us give some additional information. For every model the missing entries in Table 1.1 (Y_{ij}, $i+j > r+1$) are treated as parameters. Having obtained the MCMC output samples, it is straightforward to obtain the posterior sample for the parameters of interest $Y_{i.} = \sum_{j=r+2-i}^{r} Y_{ij}$ which express the outstanding claims of year i, by simply adding the generated values of Y_{ij}, $i+j > r+1$.

Convergence aspects were treated carefully by firstly choosing an appropriate lag interval in the MCMC output so that autocorrelation bias is minimized, and secondly by using all convergence diagnostics criteria available in CODA software (see Best *et al.* , 1995) to ensure that convergence is achieved.

1.4.3 Modelling Approaches

New modeling aspects expressed in terms of four models of the problem of outstanding claims are presented in this section. For comparison purposes, we present a Bayesian analysis of two models used in the past, the log-normal (model 1.1) and the state space model (model 1.3). We enhance these models by simultaneously modelling claim amounts and counts and using the total claim counts to specify appropriate parameter constraints. These modifications result in Models 1.2 and 1.4.

1.4.3.1 Model 1: Log-Normal Model

The simplest model for the data in Table 1.1 is a log-normal (anova-type) model. This model was investigated by Renshaw and Verrall (1994), Renshaw (1989) and Verrall (1991, 1993, 1996). Verrall (1990) produced Bayes estimates for the parameters of this model. The model is given by the formulation

$$\Upsilon_{ij} = log\frac{Y_{ij}}{inf_{ij}}, \quad \Upsilon_{ij} \sim N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = b_0 + a_i + b_j, \quad i, j = 1, \ldots, r \tag{1.1}$$

where Υ_{ij} are called log-adjusted claim amounts and $N(\mu_{ij}, \sigma^2)$ denotes the normal distribution with mean μ_{ij} and variance σ^2. In Taylor and Ashe (1983) and Verrall (1991, 1993, 1996) the alternative parametrization $\Upsilon_{ij} = log(Y_{ij}) - log(inf_{ij} \times E_i)$ with $\Upsilon_{ij} \sim N(\mu_{ij}, \sigma^2)$ is used, where E_i is a measure of exposure (for example size of portfolio for year i). This reparametrization can be easily adopted for all following models. Finally, (1.1) requires appropriate constraints to achieve identifiability of the parameters, so here we adopt the usual sum-to-zero parametrization, that is, $\sum_i a_i = \sum_j b_j = 0$. Consequently, expression (1.1) assumes that the expected log-adjusted claim amount μ_{ij} originated at year i and paid with delay of $j - 1$ years is modelled via a linear predictor which consists of the average log-adjusted claim amount b_0, a factor which reflects expected changes due to origin year a_i, and a factor depending on the delay pattern b_j.

To complete the Bayesian formulation we use the priors

$$b_0 \sim N(0, \sigma^2_{b_0}), \quad a_i \sim N(0, \sigma^2_{a_i}), \quad b_j \sim N(0, \sigma^2_{b_j}), \; i, j = 2, ..., r, \quad \tau = \sigma^{-2} \sim G(a_\tau, b_\tau)$$

with $G(a, b)$ denoting gamma distribution with mean a/b. For the kind of problems we are interested in, vague diffuse proper priors (Kass and Wasserman, 1996) are produced by using

$$\sigma^2_{b_0} = 1000, \quad \sigma^2_{a_i} = 100, \ i = 2, ..., r, \quad \sigma^2_{b_j} = 100, \ j = 2, ..., r, \quad a_\tau = b_\tau = 0.001.$$

A disadvantage of the above model is that it does not use any information from the observed counts. That is, any prediction of the missing claim amounts will be based only on the observed claim amounts. As a result, a source of information for a year (or cell) such as a sudden increase of accidents will not affect the prediction of the claim amounts.

1.4.3.2 Model 2: Log-Normal & Multinomial Model

We suggest here a two stage hierarchical model which uses both data sets in Tables 1.1 and 1.2 and the can be written, assuming $n_{ij} > 0$ for all i, j, in the two stage formulation

$$\Upsilon_{ij} = log\frac{Y_{ij}}{inf_{ij}}, \quad \Upsilon_{ij} \sim N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = b_0 + a_i + b_j + log(n_{ij}),$$

$$(n_{i1}, n_{i2}, \ldots, n_{ir})^T \sim Multinomial(p'_1, p'_2, \ldots, p'_r; T_i), \qquad log(p'_j/p'_1) = b^*_j \tag{1.2}$$

where $(n_{i1}, n_{i2}, \ldots, n_{ir})^T$ are the number of claims originated at year i and p'_j is the probability for a claim to be settled with a delay of $j-1$ years. For the first stage of the model we use as in Model 1 sum-to-zero constraints. Compared to (1.1), the linear predictor in this stage has been enhanced with the term $log(n_{ij})$. As a result, b_0 represents the average log-adjusted amount per claim finalized and a_i, b_j reflect expected differences from b_0 due to origin and delay years respectively. For the second stage we use corner constraints ($b^*_1 = 0$) to facilitate its straightforward interpretation: b^*_j represents the log-odds of an accident to be paid with a delay of $j-1$ years versus an accident paid without delay.

The second (multinomial) stage of Model 1.2 is equivalent, to the log-linear model

$$n_{ij} \sim Poisson(\lambda_{ij}), \qquad log(\lambda_{ij}) = b^*_0 + a^*_i + b^*_j$$

under the constraints $\sum_{j=1}^{r} n_{ij} = n_{i.} = T_i$, $\sum_{j=1}^{r} \lambda_{ij} = \lambda_{i.} = T_i$, where b^*_0 and a^*_i are nuisance parameters; for more details see Agresti (1990). Under the assumption that $n_{ij} > 0$ for all i, j it is precise to assume that n_{ij} follows a 'truncated at zero' $Poisson(\lambda_{ij})$. However, for the size of the data we are interested in, the above distribution is practically identical to $Poisson(\lambda_{ij})$. Had we assumed that T_i is unknown, we would have used the above log-linear model without constraints on λ_{ij} and n_{ij}. This could be useful, for example, if some kind of exposure measure is available, say the size of portfolio. Then, Model 1.2 without the constraints on λ_{ij} and n_{ij} is appropriate for predicting 'incurred but not reported claims'.

We suggest similar prior distributions as Model 1.1

$$b_0 \sim N(0, \sigma^2_{b_0}), \quad a_i \sim N(0, \sigma^2_{a_i}),\ i = 2, ..., r, \quad b_j \sim N(0, \sigma^2_{b_j}),\ j = 2, ..., r,$$

$$\tau = \sigma^{-2} \sim G(a_\tau, b_\tau), \quad b^*_j \sim N(0, \sigma^2_{b^*_j}),\ j = 2, ..., r.$$

The same values for $\sigma^2_{b_0}$, $\sigma^2_{a_i}$, $\sigma^2_{b_j}$ as in Model 1.1 can be used. For the additional parameters b^*_j we suggest $\sigma^2_{b^*_j} = 100$, for $j = 2, ..., r$.

1.4.3.3 Model 3: State Space Modelling of Claim Amounts

An alternative modelling perspective for this kind of problems is the state space (or dynamic linear) models where the parameters depend on each other in a time recursive way. A general description of MCMC in dynamic models is given by Gamerman (1998). Carter and Kohn (1994) describe how to use Gibbs sampler for general state space models and Carlin (1992) applies Gibbs sampler for state space models for actuarial time series. For application of state space models in claim amounts problem see De Jong and Zehnwirth (1983) and Verrall (1989, 1994). The state space model can be written as

$$\Upsilon_{ij} = log\frac{Y_{ij}}{inf_{ij}},\ \Upsilon_{ij} \sim N(\mu_{ij}, \sigma^2),\ \mu_{ij} = b_0 + a_i + b_{ij} \tag{1.3}$$

with the recursive associations

$$b_{ij} = b_{i-1,j} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2_\varepsilon), \quad i = 2, ..., r,$$

$$a_i = a_{i-1} + \zeta_i, \quad \zeta_i \sim N(0, \sigma^2_\zeta), \quad i = 2, ..., r$$

and corner constraints $a_1 = b_{i1} = 0, \quad i = 1, 2, \ldots, r.$

Comparing Model 3 with Model 1 we first note that b_j has been replaced by b_{ij}. Thus, the delay effect on the log-adjusted claim amounts changes with the origin year. Second, the introduced recursive associations express the belief that the parameters a_i and b_j evolve through time via known stochastic mechanisms. In fact, these mechanisms are determined by the disturbance terms ε_i and ζ_i; as σ^2_ε approaches zero (1.3) degenerates to Model 1, whereas when σ^2_ζ approaches zero the parameters a_i tend to zero. The corner constraints imply that b_0 is the expected log-adjusted claim amount for the first origin year paid without delay and a_i and b_{ij} are interpreted accordingly.

In (1.3) we only need to define prior distributions for the first state space parameters; for more details see Carlin *et al.* (1992), Carlin (1992) and Gamerman (1998). We propose priors $b_{1j} \sim N(0, \sigma^2_{b_{1j}})$ and $b_0 \sim N(0, \sigma^2_{b_0})$ with $\sigma^2_{b_{1j}} = 100$ and $\sigma^2_{b_0} = 1000$. The prior for the precision $\tau = \sigma^{-2}$ is a $G(a_\tau, b_\tau)$ density as in Model 1.1. We additionally use non-informative gamma priors for the parameters $\sigma_\varepsilon^{-2} \sim G(a_\varepsilon, b_\varepsilon)$ and $\sigma_\zeta^{-2} \sim G(a_\zeta, b_\zeta)$ with proposed values $a_\varepsilon = b_\varepsilon = a_\zeta = b_\zeta = 10^{-10}$. Finally, as in Model 1.1, we note that this model does not use any information from claim counts.

1.4.3.4 Model 4: State Space Modelling of Average Claim Amount per Accident

Here we generalise the Model 1.3 by incorporating information from data in Table 1.2. Assuming that $n_{ij} > 0$ for all i, j, we suggest

$$\Upsilon_{ij} = log\frac{Y_{ij}}{inf_{ij}}, \quad \Upsilon_{ij} \sim N(\mu_{ij}, \sigma^2), \quad \mu_{ij} = b_0 + a_i + b_{ij} + log(n_{ij}),$$

$$(n_{i1}, n_{i2}, \ldots, n_{ir})^T \sim Multinomial(p'_1, p'_2, \ldots, p'_r; T_i), \quad log(p'_j/p'_1) = b^*_j, \quad b^*_1 = 0 \qquad (1.4)$$

with the recursive associations

$$b_{ij} = b_{i-1,j} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2_\varepsilon), \quad i = 2, ..., r,$$

$$a_i = a_{i-1} + \zeta_i, \quad \zeta_i \sim N(0, \sigma^2_\zeta), \quad i = 2, ..., r.$$

In analogy with Model 2, we have added the term $log(n_{ij})$ in the linear predictor. Thus, b_0 represents the log-adjusted amount per claim finalized for the first origin year paid without delay, and a_i, b_{ij} are interpreted accordingly. The multinomial second stage formulation is interpreted exactly as in Model 2. The priors can be defined similarly as in Models 1.2 and 1.3.

1.4.4 A Real Data Example

The following data came from a major Greek insurance company. Tables 1.3-1.5 give the claim amounts, the claim counts, the total counts for car accidents and the inflation factor. Due to their nature, the main source of delay is due to claims that are notified but settled after the accident year. Liabilities that have arisen but reported later are assumed to be minimal. Moreover, the assumption of no partial payments is plausible since only a small proportion of car accident claims are paid in more than one installments.

					B			
	Year	1	2	3	4	5	6	7
	1989	527003	220645	130250	84085	72182	21656	49868
	1990	715247	341364	166001	99845	108648	91958	
	1991	1166119	428365	166410	155376	191644		
A	**1992**	1686294	647331	335290	427069			
	1993	2780948	961010	444610				
	1994	3619446	1328151					
	1995	4002087						

Table 1.3: Outstanding Claim Amounts from a Greek Insurance Company(thousands drachmas).

				B					
	Year	1	2	3	4	5	6	7	Total
	1989	6622	1943	489	138	61	223	66	9542
	1990	6943	2133	632	154	162	390		10496
	1991	8610	2216	736	651	256			12601
A	**1992**	9791	3167	1570	624				15565
	1993	11722	3192	1773					17735
	1994	13684	3664						19746
	1995	13068							18600

Table 1.4: Outstanding Claim Counts from a Greek Insurance Company.

The analysis of the data above was initiated by deinflating the data in Table 1.3 using the inflation factors in Table 1.5. Therefore, the resulting predictive amounts presented in

Year	1989	1990	1991	1992	1993	1994	1995	1996
Inflation (%)	100.0	120.4	143.9	166.6	190.6	214.2	235.6	257.0

Table 1.5: Inflation Factor for Greece.

this section should be multiplied by the corresponding inflation factor to represent amounts for a specific year (for example multiply by 257/100=2.57 to get the inflated amount for year 1996).

	Year					
Model	1990	1991	1992	1993	1994	1995
1	1107(17)	1374(22)	1904(69)	2505(118)	3026(238)	3112(556)
2	1105(19)	1322(6)	1787(48)	2400(121)	2892(271)	2950(698)
3	1103(13)	1379(27)	1896(68)	2533(167)	3013(282)	3091(475)
4	1101(3)	1330(2)	1789(8)	2433(31)	2752(59)	2767(168)

Table 1.6: Posterior Mean (Standard Deviation) for Total Claim Amounts Paid for Each Accident Year (million drachmas; adjusted for inflation).

	Year					
Model	1990	1991	1992	1993	1994	1995
1	34(17)	65(22)	215(69)	409(118)	773(238)	1413(555)
2	32(19)	13(6)	97(48)	304(121)	639(271)	1251(698)
3	30(13)	70(27)	206(68)	436(167)	760(282)	1393(475)
4	28(3)	21(2)	99(8)	336(31)	498(59)	1068(168)

Table 1.7: Posterior Mean (Standard Deviation) for Total Outstanding Claim Amounts of Each Accident Year (million drachmas; adjusted for inflation).

Posterior summaries of Models 1-4 are given in Tables 1.7 and 1.8. Note the striking difference of our proposed models 1.2 and 1.4 when compared with the existing approaches

Model	Year						Total
	1996	1997	1998	1999	2000	2001	
1	1222(338)	679(177)	470(140)	299(110)	152(59)	88(54)	2909(670)
2	1085(450)	582(215)	375(171)	191(109)	66(40)	37(29)	2336(806)
3	1166(289)	677(225)	496(227)	310(179)	161(154)	85(99)	2895(834)
4	937(136)	456(70)	353(78)	179(43)	87(24)	41(20)	2052(226)

Table 1.8: Posterior Mean (Standard Deviation) of Total Claim Amounts to be Paid in Each Future Year (million drachmas; adjusted for inflation).

expressed by Models 1.1 and 1.3 for outstanding claim amounts for 1991 and 1992. This deviation is easily explainable if we examine carefully the data in Table 1.4. The remaining outstanding claims for 1991 are only 132 and account for the 1.05% of the total claim counts (12,601). This percentage is comparably much smaller than the corresponding outstanding claim counts of 1989 and 1990 which were 3.03% and 4.48% respectively. This decrease is being taken into account by our models and the produced estimates for 1991 are appropriately adjusted.

Table 1.9 gives the posterior summaries for variance components for all models. For the data we analysed, we noticed that the state space model for claim amounts (Model 1.3) did not differ very much from the simple log-normal model. This is due to the small posterior values of σ^2_ε and may imply that no dynamic term is needed when modelling the total claim amounts. On the other hand, incorporation the claim counts (Model 4) resulted in a posterior density of σ^2_ε which gives evidence for a non-constant dynamic term. Therefore, Model 3 implies that the total payments have a similar delay pattern across years while Model 4 implies that 'payments per claim finalized' for origin year $\imath$ and delay year $\jmath$ change from year to year.

1.4.5 Discussion

In this case study we developed new models in order to analyse the well known problem of outstanding claims of insurance companies using Bayesian theory and MCMC methodology.

Posterior Value	Model Parameters σ^2 Model 1	Model 2	Model 3	Model 4	σ^2_ε Model 3	Model 4	σ^2_ζ Model 3	Model 4
mean	0.0893	0.1366	0.0623	0.00008	0.0379	0.1249	0.1091	0.0150
median	0.0816	0.1231	0.0603	0.00008	0.0002	0.1197	0.0777	0.0112
st.dev.	0.0409	0.0596	0.0399	0.00002	0.0777	0.0324	0.1227	0.0145

Table 1.9: Posterior Summaries for Model Parameters σ^2, σ^2_ε and σ^2_ζ.

The models fitted can be divided in two categories. The first category contains models that use only the information from claim amounts (Table 1.1) while the second exploits both claim amounts and counts (Tables 1.1 and 1.2). Thus the enriched family attempts to model the average payment per claim finalized or paid; this is the approach we advocate, and we believe that it improves the predictive behaviour of the model.

The models dealt with in this illustrated example can be generalised by adding other factors in the first (log-normal) stage. For example, we may assume that the variance of Υ_{ij} depends on the claim counts of the corresponding cell. Since our suggested models are already multiplicative in the error, this adjustment will improve, at least in our data, only slightly the fit.

Finally, we would like to mention that the Bayesian paradigm used in this case study did not utilize the advantage of using informative prior densities. By illustrating our results with non-informative priors, we only provide a yardstick for comparison with other approaches. However, any prior knowledges can be incorporated in our models using usual quantification arguments.

1.4.6 Full Conditional Posterior Densities of Case Study

Conditional posterior distributions needed for the MCMC implementation of the four models presented in Section 1.4.3 are given here in detail. Iterative samples from these conditional densities provide, after some burn-in period and by using an appropriate sample lag, the required samples from the posterior density.

1.4.6.1 Computations for Model 1

The model described in Section 1.4.3.1 includes parameters $b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}$. The claim amounts and counts are divided in known/observed (data) for $i+j \leq r+1$ and unknown missing (parameters) for $i+j > r+1$. Denote by $\boldsymbol{\Upsilon}^U$ the observed (inflation adjusted) log-amounts by $\boldsymbol{\Upsilon}^L$ the missing (inflation adjusted) log-amounts and by $\boldsymbol{\Upsilon}$ the matrix containing both observed and missing claim (inflation adjusted) log-amounts. Assuming that the missing data $\boldsymbol{\Upsilon}^L$ are a further set of parameters, the parameter vector is given by $(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{\Upsilon}^L)$ and the data vector is given by $(\boldsymbol{\Upsilon}^U)$. Using Bayes theorem and denoting by f the prior, conditional and marginal densities, the posterior distribution is given by

$$\begin{aligned}
f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{\Upsilon}^L | \boldsymbol{\Upsilon}^U) &\propto \\
&\propto f(\boldsymbol{\Upsilon}^U | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{\Upsilon}^L) f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{\Upsilon}^L) \\
&\propto f(\boldsymbol{\Upsilon}^U | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\boldsymbol{\Upsilon}^L | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) \\
&\propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(b_0) f(\boldsymbol{a}) f(\boldsymbol{b}) f(\sigma^{-2})
\end{aligned}$$

where $\boldsymbol{b} = (b_2, \ldots, b_r)$ and $\boldsymbol{a} = (a_2, \ldots, a_r)$.

The full conditional distributions are therefore given by

1. $f(b_0|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(b_0)$
2. $f(\boldsymbol{a}|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\boldsymbol{a})$
3. $f(\boldsymbol{b}|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\boldsymbol{b})$
4. $f(\sigma^{-2}|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\sigma^{-2})$
5. $f(\boldsymbol{\Upsilon}^L|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2})$

In the above posterior the conditional $f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2})$ is the full likelihood assuming that there are no missing data in the claim amount table; therefore,

$$f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) = (2\pi\sigma^2)^{-r^2/2} exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{r} \sum_{j=1}^{r} [\Upsilon_{ij} - b_0 - a_i - b_j)]^2 \right)$$

Thus, the resulting conditional distributions are

1.
$$f(b_0|.) = N\left(\frac{\Upsilon_{..}}{r^2+\sigma^2/\sigma^2_{b_0}}, \frac{\sigma^2}{r^2+\sigma^2/\sigma^2_{b_0}}\right), \tag{1.5}$$

where $\Upsilon_{..} = \sum_{i=1}^{r}\sum_{j=1}^{r}\Upsilon_{ij}$.

2. **[a]**
$$f(a_i|.) = N\left(\frac{\Upsilon_{i.} - \Upsilon_{1.} - r\sum_{k\neq 1,i} a_k}{2r+\sigma^2/\sigma^2_{a_i}}, \frac{\sigma^2}{2r+\sigma^2/\sigma^2_{a_i}}\right), \ i = 2,\ldots,r, \tag{1.6}$$

where $\Upsilon_{i.} = \sum_{j=1}^{r}\Upsilon_{ij}$.

[b] Set $a_1 = -\sum_{i=2}^{r} a_i$.

3. **[a]**
$$f(b_j|.) = N\left(\frac{\Upsilon_{.j} - \Upsilon_{.1} - r\sum_{k\neq 1,j} b_k}{2r+\sigma^2/\sigma^2_{b_j}}, \frac{\sigma^2}{2r+\sigma^2/\sigma^2_{b_j}}\right), \ j = 2,\ldots,r, \tag{1.7}$$

where $ln_{.j} = \sum_{i=1}^{r} log(n_{ij})$ and $\Upsilon_{.j} = \sum_{i=1}^{r}\Upsilon_{ij}$

[b] Set $b_1 = -\sum_{j=2}^{r} b_j$.

4.
$$f(\tau = \sigma^{-2}|.) = G(a_\tau + r^2/2, b_\tau + SS/2), \tag{1.8}$$

with $SS = \sum_{i=1}^{r}\sum_{j=1}^{r}(\Upsilon_{ij} - \mu_{ij})^2$ and $\mu_{ij} = b_0 + a_i + b_j$).

5.
$$f(\Upsilon_{ij}|.) = N(\mu_{ij}, \sigma^2), \quad i = 2,\ldots,r, \quad j = r-i+2,\ldots,r, \tag{1.9}$$

with $\mu_{ij} = b_0 + a_i + b_j$.

1.4.6.2 Computations for Model 2

The model introduced in 1.4.3.2 is more complicated and includes parameters $b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}$ from stage one, and $\boldsymbol{b}^*$ from stage two. Similar to above, the claim (inflation-adjusted) log-amounts and counts are divided in known/observed (data) for $i+j \leq r+1$ and unknown missing (parameters) for $i+j > r+1$. Denote by $\boldsymbol{N}^U$ and $\boldsymbol{\Upsilon}^U$ the observed claim counts and amounts, respectively, by $\boldsymbol{N}^L$ and $\boldsymbol{\Upsilon}^L$ the missing claim counts and amounts, respectively, and by $\boldsymbol{N}$ and $\boldsymbol{\Upsilon}$ the matrices containing both observed and missing claim counts and amounts, respectively. Assuming that the missing data $\boldsymbol{N}^L$ and $\boldsymbol{\Upsilon}^L$ are a further set of parameters, the parameter vector is given by $(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*, \boldsymbol{N}^L, \boldsymbol{\Upsilon}^L)$ and the data vector is given by $(\boldsymbol{N}^U, \boldsymbol{\Upsilon}^U)$. Using Bayes theorem and denoting by f the prior, conditional and marginal densities, the posterior distribution is given by

$$
\begin{aligned}
f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*, \boldsymbol{N}^L, \boldsymbol{\Upsilon}^L | \boldsymbol{N}^U, \boldsymbol{\Upsilon}^U) \;\; \propto & \\
& \propto \; f(\boldsymbol{N}^U, \boldsymbol{\Upsilon}^U | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*, \boldsymbol{N}^L, \boldsymbol{\Upsilon}^L) f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*, \boldsymbol{N}^L, \boldsymbol{\Upsilon}^L) \\
& \propto \; f(\boldsymbol{N}^U, \boldsymbol{\Upsilon}^U | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*) f(\boldsymbol{N}^L, \boldsymbol{\Upsilon}^L | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*) f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*) \\
& \propto \; f(\boldsymbol{N}, \boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{b}^*) f(b_0) f(\boldsymbol{a}) f(\boldsymbol{b}) f(\sigma^{-2}) f(\boldsymbol{b}^*) \\
& \propto \; f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) f(\boldsymbol{N} | \boldsymbol{b}^*) f(b_0) f(\boldsymbol{a}) f(\boldsymbol{b}) f(\sigma^{-2}) f(\boldsymbol{b}^*)
\end{aligned}
$$

where $\boldsymbol{b} = (b_2, \ldots, b_r)$, $\boldsymbol{a} = (a_2, \ldots, a_r)$ and $\boldsymbol{b}^* = (b_2^*, \ldots, b_r^*)$.

The full conditional distributions are therefore given by

1. $f(b_0|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) f(b_0)$
2. $f(\boldsymbol{a}|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) f(\boldsymbol{a})$
3. $f(\boldsymbol{b}|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) f(\boldsymbol{b})$
4. $f(\sigma^{-2}|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) f(\sigma^{-2})$
5. $f(\boldsymbol{\Upsilon}^L|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N})$
6. $f(\boldsymbol{b}^*|.) \propto f(\boldsymbol{N} | \boldsymbol{b}^*) f(\boldsymbol{b}^*)$
7. $f(\boldsymbol{N}^L|.) \propto f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) f(\boldsymbol{N} | \boldsymbol{b}^*)$

In the above posterior the conditional $f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N})$ is the full likelihood for the first stage assuming that there are no missing data in the claim amount table; therefore,

$$
f(\boldsymbol{\Upsilon} | b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \boldsymbol{N}) = (2\pi\sigma^2)^{-r^2/2} exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{r} \sum_{j=1}^{r} [\Upsilon_{ij} - b_0 - a_i - b_j - log(n_{ij})]^2 \right)
$$

The full likelihood $f(\boldsymbol{N} | \boldsymbol{b}^*)$ of the second stage, assuming no missing claim counts, can be written as

$$
f(\boldsymbol{N} | \boldsymbol{b}^*) = exp\left(\sum_{i=1}^{r} log(T_i!) - \sum_{i=1}^{r} \sum_{j=1}^{r} log(n_{ij}!) + \sum_{j=2}^{r} n_{.j} b_j^* - n_{..} log(\sum_{k=1}^{r} e^{b_k^*}) \right),
$$

where $n_{..} = \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij} = \sum_{i=1}^{r} T_i$ and $n_{.j} = \sum_{i=1}^{r} n_{ij}$. Thus, the resulting conditional distributions are

1. $$f(b_0|.) = N\left(\frac{\Upsilon_{..} - ln_{..}}{r^2 + \sigma^2/\sigma^2_{b_0}}, \frac{\sigma^2}{r^2 + \sigma^2/\sigma^2_{b_0}}\right), \quad (1.10)$$

   where $ln_{..} = \sum_{i=1}^{r}\sum_{j=1}^{r} log(n_{ij})$ and $\Upsilon_{..} = \sum_{i=1}^{r}\sum_{j=1}^{r} \Upsilon_{ij}$.

2. **[a]**
   $$f(a_i|.) = N\left(\frac{\Upsilon_{i.} - \Upsilon_{1.} - (ln_{i.} - ln_{1.}) - r\sum_{k\neq 1,i} a_k}{2r + \sigma^2/\sigma^2_{a_i}}, \frac{\sigma^2}{2r + \sigma^2/\sigma^2_{a_i}}\right), \; i = 2, \ldots, r, \quad (1.11)$$

   where $ln_{i.} = \sum_{j=1}^{r} log(n_{ij})$ and $\Upsilon_{i.} = \sum_{j=1}^{r} \Upsilon_{ij}$.

   [b] Set $a_1 = -\sum_{i=2}^{r} a_i$.

3. **[a]**
   $$f(b_j|.) = N\left(\frac{\Upsilon_{.j} - \Upsilon_{.1} - (ln_{.j} - ln_{.1}) - r\sum_{k\neq 1,j} b_k}{2r + \sigma^2/\sigma^2_{b_j}}, \frac{\sigma^2}{2r + \sigma^2/\sigma^2_{b_j}}\right), \; j = 2, \ldots, r, \quad (1.12)$$

   where $ln_{.j} = \sum_{i=1}^{r} log(n_{ij})$ and $\Upsilon_{.j} = \sum_{i=1}^{r} \Upsilon_{ij}$

   [b] Set $b_1 = -\sum_{j=2}^{r} b_j$.

4. $f(\tau = \sigma^{-2}|.)$ is given by (1.8) with $\mu_{ij} = b_0 + a_i + b_j + log(n_{ij})$.

5. $f(\Upsilon_{ij}|.)$ is given by (1.9) with $\mu_{ij} = b_0 + a_i + b_j + log(n_{ij})$.

6. **[a]**
   $$f(b^*_j|.) \propto exp\left(b^*_j n_{.j} - n_{..}log(\sum_{k=1}^{r} e^{b^*_k}) - 0.5 b^{*2}_j/\sigma^2_{b^*_j}\right), \; j = 2, \ldots, r, \quad (1.13)$$

   where $n_{.j} = \sum_{i=1}^{r} n_{ij}$.

   [b] Set $b^*_1 = 0$.

   To obtain a sample from (1.13) we may use either Metropolis-Hastings algorithm or Gilks and Wild (1992) adaptive rejection sampling for log-concave distributions. Both methods provide similar convergence rates.

7. The full conditional posterior of the missing counts n_{ij} for $j = r-i+2,\dots,r-1$, $i = 3,\dots,r$ is complicated since

$$f(n_{ij}|.) \propto f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{N}) f(\boldsymbol{N}|\boldsymbol{b}^*)$$

The constraint $T_i = \sum_{j=1}^{r} n_{ij}$ reduces the above posterior to

$$f(n_{ij}|.) \propto f(\Upsilon_{ij}|b_0, a_i, b_j, n_{ij}, \Upsilon_{ir}) f(\Upsilon_{ir}|b_0, a_i, b_r, n_{ir}) f(n_{ij}|b_j^*, n_{.r}) f(n_{ir}|b_r^*)$$

where $n_{ir} = T_i - \sum_{j=1}^{r-1} n_{ij}$. Therefore,

$$f(n_{ij}|.) \propto \frac{[p'_j]^{n_{ij}}}{n_{ij}!} \frac{[p'_r]^{\Omega_{ij}-n_{ij}}}{(\Omega_{ij}-n_{ij})!} \exp\left(\Psi_{ij}(n_{ij}) + \Psi_{ir}(\Omega_{ij}-n_{ij})\right) \tag{1.14}$$

where $\Omega_{ij} = T_i - \sum_{k\neq j,r} n_{ik}$ and $\Psi_{ij}(n_{ij}) = -\frac{1}{2\sigma^2}(\Upsilon_{ij} - b_0 - a_i - b_j - log(n_{ij}))^2$.

We sample from $f(n_{ij}|.)$ by using the following Metropolis-Hastings step. Propose missing n'_{ij} and $n'_{ir} = \Omega_{ij} - n'_{ij}$, with $i = 3,\dots,r$ and $j = r-i+2,\dots,r-1$ from

$$Binomial(p''_j, \Omega_{ij}),\ p''_j = \frac{p'_j}{p'_j + p'_r} = (1 + exp(b_r^* - b_j^*))^{-1}.$$

Accept the proposed move with probability

$$a = min\{1, exp(\Psi_{ij}(n'_{ij}) + \Psi_{ir}(\Omega_{ij} - n'_{ij}) - \Psi_{ij}(n_{ij}) - \Psi_{ir}(\Omega_{ij} - n_{ij}))\}. \tag{1.15}$$

1.4.6.3 Computations for Model 3

The dynamic model described in Section 1.4.3.3 is an extension of model of Section 1.4.3.1 includes parameters $b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \sigma_\varepsilon^{-2}, \sigma_\zeta^{-2}$, where $\boldsymbol{b} = (b_{12},\dots,b_{1r}, b_{22},\dots,b_{2r},\dots,b_{rr})$. Using Bayes theorem the posterior distribution is given by

$$f(b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}, \sigma_\epsilon^{-2}, \sigma_\zeta^{-2}, \boldsymbol{\Upsilon}^L|\boldsymbol{\Upsilon}^U) \quad \propto$$

$$\propto \quad f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(b_0) f(\boldsymbol{a}|\sigma_\zeta^{-2}) f(\boldsymbol{b}|\sigma_\varepsilon^{-2}) f(\sigma^{-2}) f(\sigma_\varepsilon^{-2}) f(\sigma_\zeta^{-2}).$$

The full conditional distributions are therefore given by

1. $f(b_0|.) \propto f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(b_0)$
2. $f(\boldsymbol{a}|.) \propto f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\boldsymbol{a}|\sigma_\zeta^2)$

3. $f(\boldsymbol{b}|.) \propto f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\boldsymbol{b}|\sigma_\varepsilon^2)$

4. $f(\sigma^{-2}|.) \propto f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) f(\sigma^{-2})$

5. $f(\sigma_\varepsilon^{-2}|.) \propto f(\boldsymbol{b}|\sigma_\varepsilon^{-2}) f(\sigma_\varepsilon^{-2})$

6. $f(\sigma_\zeta^{-2}|.) \propto f(\boldsymbol{a}|\sigma_\zeta^{-2}) f(\sigma_\zeta^{-2})$

7. $f(\boldsymbol{\Upsilon}^L|.) \propto f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2})$

Similar to Model 1, the conditional $f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2})$ is the full likelihood assuming that there are no missing data in the claim amount table; therefore,

$$f(\boldsymbol{\Upsilon}|b_0, \boldsymbol{a}, \boldsymbol{b}, \sigma^{-2}) = (2\pi\sigma^2)^{-r^2/2} exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{r} \sum_{j=1}^{r} [\Upsilon_{ij} - b_0 - a_i - b_{ij})]^2 \right)$$

Thus, the resulting conditional distributions are

1.
$$f(b_0|.) = N\left(\frac{\Upsilon_{..} - r^2\bar{a} - r^2\bar{b}}{r^2 + \sigma^2/\sigma_{b_0}^2}, \frac{\sigma^2}{r^2 + \sigma^2/\sigma_{b_0}^2} \right), \tag{1.16}$$
where $\bar{a} = r^{-1}\sum_i a_i$ and $\bar{b} = r^{-2}\sum_{ij} b_{ij}$.

2. **[a]** Set $a_1 = 0$.

[b]
$$f(a_i|.) = N\left(\frac{\Upsilon_{i.} - rb_0 - b_{i.} + (a_{i+1} + a_{i-1})\sigma^2/\sigma_\zeta^2}{r + 2\sigma^2/\sigma_\zeta^2}, \frac{\sigma^2}{r + 2\sigma^2/\sigma_\zeta^2} \right), \; i = 2, \ldots, r-1 \tag{1.17}$$

where $b_{i.} = \sum_j b_{ij}$.

[c]
$$f(a_r|.) = N\left(\frac{\Upsilon_{r.} - rb_0 - b_{r.} + a_{r-1}\sigma^2/\sigma_\zeta^2}{r + \sigma^2/\sigma_\zeta^2}, \frac{\sigma^2}{r + \sigma^2/\sigma_\zeta^2} \right). \tag{1.18}$$

3. **[a]** Set $b_{i1} = 0$ for $i = 1, \ldots, r$.

[b]
$$f(b_{1j}|.) = N\left(\frac{\Upsilon_{1j} - b_0 + b_{2j}\sigma^2/\sigma_\varepsilon^2}{1 + \sigma^2/\sigma_\varepsilon^2 + \sigma^2/\sigma_{b_{1j}}^2}, \frac{\sigma^2}{1 + \sigma^2/\sigma_\varepsilon^2 + \sigma^2/\sigma_{b_{1j}}^2} \right), \; j = 2, \ldots, r. \tag{1.19}$$

[c]
$$f(b_{ij}|.) = N\left(\frac{\Upsilon_{ij} - b_0 - a_i + (b_{i-1,j} + b_{i+1,j})\sigma^2/\sigma_\varepsilon^2}{1 + 2\sigma^2/\sigma_\varepsilon^2}, \frac{\sigma^2}{1 + 2\sigma^2/\sigma_\varepsilon^2} \right), \tag{1.20}$$

for $i = 2, \ldots, r-1, \; j = 2, \ldots, r$.

[d]

$$f(b_{rj}|.) = N\left(\frac{\Upsilon_{rj} - b_0 - a_r + b_{r-1,j}\sigma^2/\sigma_\varepsilon^2}{1+\sigma^2/\sigma_\varepsilon^2}, \frac{\sigma^2}{1+\sigma^2/\sigma_\varepsilon^2}\right), \; j = 2, \ldots, r. \tag{1.21}$$

4. $f(\tau|.)$ is given by equation (1.8), using $\mu_{ij} = b_0 + a_i + b_{ij}$.

5. $f(\Upsilon_{ij}|.)$, for $i + j > r + 1$, is given by equation (1.9) using $\mu_{ij} = b_0 + a_i + b_{ij}$.

6. $$f(\sigma_\varepsilon^{-2}|.) = G\left(a_\varepsilon + (r-1)^2/2, \; b_\varepsilon + \sum_{i=2}^{r}\sum_{j=2}^{r}(b_{ij} - b_{i-1,j})^2/2\right). \tag{1.22}$$

7. $$f(\sigma_\zeta^{-2}|.) = G\left(a_\zeta + (r-1)/2, \; b_\zeta + \sum_{i=2}^{r}(a_i - a_{i-1})^2/2\right). \tag{1.23}$$

1.4.6.4 Computations for Model 4

The first stage of Model 4 is similar to model 3 but we substitute Υ_{ij} by $\Upsilon_{ij}^* = log[Y_{ij}]$ - $log[n_{ij} inf_{ij}]$ in all conditional distributions (1.16 - 1.21). The stage two is equivalent to the second stage of Model 2. In more detail we have

1. $f(b_0|.)$ is given by (1.16) if we substitute Υ_{ij} by Υ_{ij}^*.

2. **[a]** Set $a_1 = 0$.

   [b] $f(a_i|.)$ for $i = 2, \ldots, r-1$ is given by (1.17) if we substitute Υ_{ij} by Υ_{ij}^*.

   [c] $f(a_r|.)$ is given by (1.18) if we substitute Υ_{ij} by Υ_{ij}^*.

3. **[a]** Set $b_{i1} = 0$ for $i = 1, \ldots, r$.

   [b] $f(b_{1j}|.)$ for $j = 2, \ldots, r$ is given by (1.19) if we substitute Υ_{ij} by Υ_{ij}^*.

   [c] $f(b_{ij}|.)$ for $i = 2, \ldots, r-1, \; j = 2, \ldots, r$ is given by (1.20) if we substitute Υ_{ij} by Υ_{ij}^*.

   [d] $f(b_{rj}|.)$ for $j = 2, \ldots, r$ is given by (1.21) if we substitute Υ_{ij} by Υ_{ij}^*.

4. $f(\tau|.)$ is given by equation (1.8), using $\mu_{ij} = b_0 + a_i + b_{ij} - log(n_{ij})$.

5. $f(\Upsilon_{ij}|.)$, for $i+j > r+1$, is given by equation (1.9) using $\mu_{ij} = b_0 + a_i + b_{ij} - log(n_{ij})$.

6. $f(\sigma_\varepsilon^{-2}|.)$ is given by (1.22).

7. $f(\sigma_\zeta^{-2}|.)$ is given by (1.23).

8. $f(b_j^*|.)$ is given by (1.13).

9. The full conditional posterior of the missing counts $f(n_{ij}|.)$ for $j = r - i + 2, \ldots, r - 1$, $i = 3, \ldots, r$ is given by (1.14) with $\Psi_{ij}(n_{ij}) = -0.5\sigma^{-2}[\Upsilon_{ij} - b_0 - a_i - b_{ij} - log(n_{ij})]^2$.

   In order to achieve to achieve an optimal acceptance rate we propose a simultaneous updating scheme of n_{ij}, b_{ij} and b_{ir} when $i = 3, \ldots, r$ and $j = r - i + 2, \ldots, r$. The corresponding joint full conditional posterior of these parameters is given by given by an equation of type (1.14) substituting Ψ_{ij} with $\Psi^*_{ij}(n_{ij}, b_{ij}) = -0.5\sigma^{-2}[\Upsilon_{ij} - b_0 - a_i - b_{ij} - log(n_{ij})]^2 - 0.5\sigma_\varepsilon^{-2}[(b_{ij} - b_{i-1,j})^2 + (b_{i+1,j} - b_{i,j})^2]$.

   We used the following metropolis step. We propose candidate $n'_{ij}, b'_{ij}, n'_{ir}$ from the proposal densities

$$q(n'_{ij}, b'_{ij}, b'_{ir}|n_{ij}, b_{ij}, b_{ir}) = q(n'_{ij}|n_{ij}, b_{ij}, b_{ir})q(b'_{ij}|n'_{ij}, n_{ij}, b_{ij}, b_{ir})q(b'_{ir}|n'_{ij}, n_{ij}, b_{ij}, b_{ir})$$

   with

$$\begin{aligned} q(n'_{ij}|n_{ij}, b_{ij}, b_{ir}) &= Binomial([1 + exp(b_r^* - b_j^*)]^{-1}, \Omega_{ij}) \\ q(b'_{ij}|n'_{ij}, n_{ij}, b_{ij}, b_{ir}) &= N(b_{ij} + log(n_{ij}) - log(n'_{ij}), \bar{\sigma}^2_{ij}), \\ q(b'_{ir}|n'_{ij}, n_{ij}, b_{ij}, b_{ir}) &= N(b_{ir} + log(\Omega_{ij} - n_{ij}) - log(\Omega_{ij} - n'_{ij}), \bar{\sigma}^2_{ir}) \end{aligned}$$

   where $\bar{\sigma}^2_{ij}$ and $\bar{\sigma}^2_{ir}$ are metropolis parameters that should be calibrated appropriately to achieve a desired acceptance rate.

   Accept the proposed move with probability

$$\alpha = min\left\{1, \frac{exp[\Psi^*_{ij}(n'_{ij}, b'_{ij}) + \Psi^*_{ir}(\Omega_{ij} - n'_{ij}, b'_{ir})]q(b_{ij}, b_{ir}|n_{ij}, n'_{ij}, b'_{ij}, b'_{ir})}{exp[\Psi^*_{ij}(n_{ij}, b_{ij}) + \Psi^*_{ir}(\Omega_{ij} - n_{ij}, b_{ir})]q(b'_{ij}, b'_{ir}|n'_{ij}, n_{ij}, b_{ij}, b_{ir})}\right\}. \quad (1.24)$$

Chapter 2

Model Selection Strategies

Statistical models are used for two important reasons: interpretation of casual relationships between certain characteristics of the population (for example relationship of cancer and smoking) and prediction of future outcomes (for example price of a product).

A complete model formulation includes specification of a response variable and the covariates, the connection between them, the distributional form of the response (and of the covariates) or any other characteristic needed. Model selection is any procedure that determines the exact form of the structure of a model.

The selection of the final model is made using certain scientific procedures that evaluate the performance of each model and select the 'best' one. The most popular of these methods are presented in this chapter giving more weight to Bayesian model selection.

2.1 Stepwise Strategies Using Significance Tests

Classical model selection procedures involve sequential comparisons with significance tests. For example, in Generalised Linear Models (GLM) these tests are based on the F or χ^2 distributions. Since the number of competing models may be large (for example, 15 regressors in any generalised linear model result in 32768 models) statisticians have constructed 'clever' and 'computationally cheap' alternative methods. The most popular methods are stepwise strategies based on the original idea of Efroymson (1960). The most common stepwise strategies are the forward selection and the backward elimination (see, for example, Efroymson, 1966, Draper and Smith, 1966). Natural hybrids of these methods are stepwise backward or stepwise forward procedures.

Generally, for any generalised linear model type setup, the forward selection is defined as following: we initially consider the model without any variable in the model equation (constant model) and continue by adding in each step the variable with the smallest p-value until a stopping rule is satisfied. This stopping rule is usually of the type p-value$>p^*$ or $F<F_{in}$; where p^* and F_{in} are arbitrary significance values. Usually $p^*=0.05$ and F_{in} is the 95th percentile of the corresponding F distribution. On the other hand, backward elimination starts from the model with all variables in the model equation (full model) and continues by removing in each step the variable with the largest p-value until a similar stopping rule is satisfied (for example p-value$<p^*$ or $F>F_{out}$); where F_{out} is again the $(1-p^*)$th quantile of the corresponding F distribution. Stepwise backward (or stepwise forward) strategy is similar to the above methods but it also considers in each step whether previously excluded variables should be included (or whether previously included variables should be excluded). Comprehensive description of stepwise procedures is provided by Thompson (1978). Stepwise methods became very popular but according to Hocking (1976)

> *stepwise procedures have been criticized on many counts, the most common being that neither FS, BE or ES [forward, backward or stepwise forward strategies] will assure, with the obvious exceptions, that the "best" subset of a given size will be revealed.*

while Miller (1984) argues that 'none of these "cheap" methods guarantees to find the best fitting subsets'.

Raftery (1995) reports that in social sciences p-values and significance tests are less and less used for hypothesis testing and model selection. The main reason is that social researchers have extremely large datasets that report small p-values even if the hypothesized model is plausible and 'inspection of data fails to reveal any striking discrepancies with it'. Moreover, Freedman (1983) claims that the interpretation of p-values is not the same in all model comparisons and exact significance level cannot be calculated since stepwise methods are sequential application of simple significance tests. Moreover, Miller (1984) notes that the maximum F-to-enter statistic[1] 'is not even remotely like an F-distribution'. Further

[1] F-to-enter statistic: the F-statistic used to test whether an excluded term should be included in the current model.

criticism on the inconsistent use of p-values as measures of evidence are given by Berger and Selke (1987), Berger and Delampady (1987), Delampady and Berger (1990), Schervish (1996) and Bayarri and Berger (1998a,b,c).

The major drawback of classical methods is the selection of a single model. Consider the case where many models are plausible and sufficiently fit the data but have totally different interpretation. Choosing a single model does not account for model uncertainty and therefore inferences, conditionally on a single selected model, may be biased. The notion of selecting more than one model is not new since Gorman and Toman (1966) argued in favour of this idea. Moreover, Hocking (1976) supported this idea by noting that 'It is unlikely that there is a single best subset but rather several equal good ones'. Miller (1984) also recommends that 'the best 10 or 20 subsets of each size, not just the best one, should be saved'. Finally, Chatfield (1995) argues that 'model uncertainty is present' and that statisticians should stop pretending that does not exists.

Although some significance tests for comparing non-nested models have recently been developed, see for example Panaretos *et al.* (1997), generally classical procedures can only compare nested models; one model m_1 is nested to another model m_2 if all the terms of model m_1 are also included in model m_2. Also note that the number of models considered may be large and therefore stepwise procedures cannot explore the whole model space.

Additional disadvantages are reported by Volinsky *et al.* (1996, 1997) and references therein. They include the selection of explanatory variables that are pure noise; the selection of different models by different stepwise methods; the phenomenon reported by Hocking (1976) and Miller (1984) where a variable that is firstly added in forward selection may be firstly removed in backward elimination; and the argument that the usual 5% significance level is arbitrary selected and does not reflect the real significance level in stepwise procedures. The last drawback was examined by Bendel and Afifi (1977) reporting that a value of 0.15 significance level for forward selection results in selection of plausible models while Kennedy and Bancroft (1975) recommended significance level equal to 0.25 for forward selection and 0.10 for backward elimination.

Finally, the efficiency of stepwise procedures can be summarised by the words of Copas (1984) arguing that stepwise methods are 'frequently used', 'frequently abused' and 'little understood' methods of applied statistics.

2.2 Bayesian Model Selection Techniques

Bayesian model selection is mainly based on the calculation of posterior model probabilities and posterior odds. Bayes factor is the posterior odds when all alternative models have equal prior probability. A massive work on Bayes factors and their applications has been published. Some well distinctive publications are provided by Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982), Kass (1993), Kass and Raftery (1995), Hoeting *et al.* (1998) and Wasserman (1997). For applications see Kass and Raftery (1995) and references therein. The evaluation of the integrals involved in posterior odds is the main topic for many publications. Analytic or Monte Carlo approximations where initially used; see Gelfand and Dey (1994), Kass and Raftery (1995), Verdinelli and Wasserman (1995), Kass and Wasserman (1995), Chib (1995), Raftery (1996a,b), DiCiccio *et al.* (1997) and Pauler (1998). Recent computational advances led to the construction of highly complicated MCMC algorithms for the computation of posterior probabilities. MCMC methods for model selection are constructed by George and McCulloch (1993), Madigan and Raftery (1994), Carlin and Chib (1995), Green (1995), Clyde *et al.* (1996), Hoeting *et al.* (1996), Smith and Kohn (1996), Raftery *et al.* (1997), and Kuo and Mallick (1998).

Lindley (1957) and Bartlett (1957) paradoxes led Bayesians to search for new 'improved' model selection measures. These approaches include Bayes factor variants such as the posterior Bayes factor by Aitkin (1991), the fractional Bayes factor by O'Hagan (1995) and the intrinsic Bayes factor by Berger and Pericchi (1996a,b).

Other approaches include predictive criteria introduced by Ibrahim and Laud (1994) and Bayesian predictive p-values; see Guttman, 1967, Box, 1980, Rubin, 1984, Meng, 1994, Bayarri and Berger, 1998a,b,c.

2.2.1 Bayesian Model Comparison

2.2.1.1 Definition of Posterior Probabilities, Odds and Bayes Factor

Bayesian model selection and hypothesis testing are based on posterior probabilities, on posterior odds and on Bayes factors. Kass and Raftery (1995) argue that Bayesian methods can 'evaluate the evidence in favour of the null hypothesis', compare two or more non-nested models, draw inferences without ignoring model uncertainty and finally determine which set of explanatory variables give better predictive results.

Consider two competing models m_0 and m_1 and suppose that the data $\boldsymbol{y}$ are considered to have been generated by one of these two models. Each model $m \in \{m_0, m_1\}$ specifies the distribution of $\boldsymbol{Y}$, $f(\boldsymbol{y}|m, \boldsymbol{\beta}_{(m)})$ apart from an unknown parameter vector $\boldsymbol{\beta}_{(m)} \in \mathcal{B}_m$, where $\mathcal{B}_m$ is the set of all possible values for the coefficients of model m. If $f(m)$ is the prior probability of model m, then, using the Bayes theorem, the posterior probability for a model is given by

$$f(m|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|m)f(m)}{f(\boldsymbol{y}|m_0)f(m_0) + f(\boldsymbol{y}|m_1)f(m_1)}$$

where $m \in \{m_0, m_1\}$ and $f(m_0) + f(m_1) = 1$. The posterior odds PO_{01} of model m_0 versus model m_1 is given by

$$PO_{01} = \frac{f(m_0|\boldsymbol{y})}{f(m_1|\boldsymbol{y})} = \frac{f(\boldsymbol{y}|m_0)}{f(\boldsymbol{y}|m_1)} \times \frac{f(m_0)}{f(m_1)}.$$

The quantity

$$B_{01} = \frac{f(\boldsymbol{y}|m_0)}{f(\boldsymbol{y}|m_1)}$$

is called *Bayes factor* of model m_0 against model m_1. According to Kass and Raftery (1995), the quantity $f(\boldsymbol{y}|m)$ is '*the predictive probability of the data*' under model m, that is the probability to get the actually observed data before any data were available under the assumption that model m holds. This predictive probability is given by

$$f(\boldsymbol{y}|m) = \int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|m) d\boldsymbol{\beta}_{(m)} \tag{2.1}$$

where $f(\boldsymbol{\beta}_{(m)}|m)$ is the conditional prior distribution of $\boldsymbol{\beta}_{(m)}$, the model parameters for model m. From the above we have

$$\textbf{Posterior odds} = \textbf{Bayes factor} \times \textbf{prior odds}.$$

The above model comparison can be extended for more than two competing models. Consider the set of models $\mathcal{M} = \{m_1, m_2, \cdots, m_{|\mathcal{M}|}\}$ then the posterior probability is given by

$$f(m|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|m)f(m)}{\sum\limits_{m_l \in \mathcal{M}} f(\boldsymbol{y}|m_l)f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l, m} \right)^{-1}, \quad m \in \mathcal{M}, \tag{2.2}$$

where $\mathcal{M}$ and $|\mathcal{M}|$ denote the set and the number of models under consideration.

Possible interpretations of Bayes factor are given by Tables 2.1 and 2.2 provided by Kass and Raftery (1995). From the above we can easily conclude that posterior odds and Bayes

$log_{10}(B_{10})$	B_{10}	Evidence against H_0
0.0 to 0.5	1.0 to 3.2	Not worth than a bare mention
0.5 to 1.0	3.2 to 10	Substantial
1.0 to 2.0	10 to 100	Strong
greater than 2	greater than 100	Decisive

Table 2.1: Bayes Factor Interpretation according to Kass and Raftery (log of 10).

$ln(B_{10})$	B_{10}	Evidence against H_0
0 to 2	1 to 3	Not worth than a bare mention
2 to 5	3 to 12	Positive
5 to 10	12 to 150	Strong
greater than 10	greater than 150	Decisive

Table 2.2: Bayes Factor Interpretation according to Kass and Raftery (Natural logarithm).

factors are invariant to any set of competing models $\mathcal{M}$ used, while posterior probabilities are not.

Bayes factor of model m_1 against m_0, B_{10}, evaluates the evidence *against* the null hypothesis which is familiar to classical significance tests. On the other hand, the Bayes factor B_{01} evaluates the evidence *in favour* of the null hypothesis which is not feasible in classical significance tests.

This integral involved in (2.1) is analytically tractable only in certain restricted examples and therefore asymptotic approximations or Monte Carlo methods are used instead; for more details see Sections 2.2.1.2 and 2.2.1.3.

2.2.1.2 Analytic Approximations of Bayes Factor

The most popular approximation is the Laplace approximation used by Tierney and Kadane (1986), Tierney *et al.* (1989) and Erkanli (1994) resulting in

$$f(\boldsymbol{y}|m) \approx (2\pi)^{d(m)/2} \left| \boldsymbol{\mathcal{I}}_{\boldsymbol{\beta}_{(m)}} \right|^{\frac{1}{2}} f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m)}, m) f(\breve{\boldsymbol{\beta}}_{(m)}|m) \qquad (2.3)$$

where $d(m)$ is the dimension of model m, $\boldsymbol{\mathcal{I}}_{\boldsymbol{\beta}_{(m)}}$ is the Hessian matrix of second derivatives of the log-posterior distribution and $\breve{\boldsymbol{\beta}}_{(m)}$ is the posterior mode for model m. A simpler but less accurate variant can be obtained by substituting $\boldsymbol{\mathcal{I}}_{\boldsymbol{\beta}_{(m)}}$ and $\breve{\boldsymbol{\beta}}_{(m)}$ by $\hat{\boldsymbol{\Sigma}}_{(m)}$ and $\hat{\boldsymbol{\beta}}_{(m)}$ which are the observed information matrix and maximum likelihood estimates of model m, respectively. The last approximation can easily be calculated from any standard statistical software that provides the maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{(m)}$, the observed information matrix $\hat{\boldsymbol{\Sigma}}_{(m)}$ and the value of the maximized likelihood, $f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m)}, m)$. A variant of the approximation is the Laplace-Metropolis estimator at which we use the posterior mode (or sub-optimally, the posterior median) and covariance matrix estimated by a sample generated from the posterior distribution; for details see Raftery (1996b).

Another common and simple approximation is based on Schwarz (1978) criterion given by

$$S = log(f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m_1)}, m_1) - log(f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m_0)}, m_0) - \frac{1}{2}[d(m_1) - d(m_0)]log(n)$$

where n is the sample size. As sample size n some statisticians argue that we should use the dimension of $\boldsymbol{y}$ vector while Raftery (1996a) defines it as the dimension of $\boldsymbol{y}$ in normal models, as the sum of all Bernoulli trials in binomial models and as the sum of all counts in Poisson models. The main advantage of the above statistic is its independence of any prior distribution. Due to its property that

$$\frac{S - log(B_{10})}{log(B_{10})} \rightarrow 0 \quad \text{when} \quad n \rightarrow \infty$$

it can be used as an approximation of the logarithm of the Bayes factor. Moreover, the quantity

$$\textbf{BIC} = \mathbf{-2} \times \textbf{Schwarz Criterion}$$

is called *Bayes Information Criterion* and is provided by many statistical packages. From the above we also have that

$$-2logB_{01} \rightarrow -2log(LR_{01}) + \{d(m_0) - d(m_1)\}log(n) \quad \text{when} \quad n \rightarrow \infty \qquad (2.4)$$

where $-2log(LR_{01}) = -2log[f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m_0)}, m_0)] + 2log[f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m_1)}, m_1)]$ is the deviance measure (see McCullagh and Nelder, 1983). Calculation of all Bayes factors against the full model leads immediately to posterior model probabilities.

For further details on Schwarz approximation see Kass and Wasserman (1995), Raftery (1996a) and Pauler (1998).

Berger and Pericchi (1998) underline that both Laplace and BIC approximations are 'valid only for "nice" problems'. The use of these approximations is limited to large sample sizes, models with regular asymptotics and models for which the likelihood is not concentrated in the boundary of the parameter space (for example in one-sided tests).

2.2.1.3 Monte Carlo Estimates of Bayes Factor

Monte Carlo methods can be used to calculate $f(\boldsymbol{y}|m)$. A simple Monte Carlo method is applied by generating a sample $\{\boldsymbol{\beta}^{(1)}_{(m)}, \boldsymbol{\beta}^{(2)}_{(m)}, \cdots, \boldsymbol{\beta}^{(t)}_{(m)}\}$ from the prior distribution $f(\boldsymbol{\beta}_{(m)}|m)$ and calculate

$$\hat{f}(\boldsymbol{y}|m) = \frac{1}{t}\sum_{t'=1}^{t} f(\boldsymbol{y}|\boldsymbol{\beta}^{(t')}_{(m)})$$

which is the average of the likelihood for each of the sampled values $\boldsymbol{\beta}^{(t')}_{(m)}$. The above estimator is unstable when the prior is diffuse or the likelihood is much more concentrated than the prior. In such cases the simulation will be inefficient since most of the simulated values will have low likelihood values and therefore the estimates will be dominated by few large values. Moreover, the variance of the above estimator will be large and the convergence of the estimator to its true value very slow.

A more precise Monte Carlo estimator is provided by *importance sampling*. This method involves simulation of $\boldsymbol{\beta}^{(t')}_{(m)}$ from an arbitrary density $g^*(\boldsymbol{\beta}_{(m)}|m)$ and estimate $f(\boldsymbol{y}|m)$ by

$$\hat{f}(\boldsymbol{y}|m) = \frac{\sum_{t'=1}^{t} w^*_{t'} f(\boldsymbol{y}|\boldsymbol{\beta}^{(t')}_{(m)}, m)}{\sum_{t'=1}^{t} w^*_{t'}}, \tag{2.5}$$

where $w^*_{t'} = f(\boldsymbol{\beta}^{(t')}_{(m)}|m)/g^*(\boldsymbol{\beta}^{(t')}_{(m)}|m)$. If we use as $g^*(\boldsymbol{\beta}_{(m)}|m)$ the posterior distribution $f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})$ then the estimator is given by

$$\hat{f}(\boldsymbol{y}|m) = \left\{\frac{1}{t}\sum_{t'=1}^{t} \left[f(\boldsymbol{y}|\boldsymbol{\beta}^{(t')}_{(m)}, m)\right]^{-1}\right\}^{-1}.$$

Details for Monte Carlo estimates are given by Kass and Raftery (1995) and references therein. More sophisticated Monte Carlo estimates are supplied by Gelfand and Dey (1994), Newton and Raftery (1994), Chib (1995) and DiCiccio *et al.* (1997).

2.2.1.4 Interpretation of Prior and Posterior Model Probabilities

The interpretation of both prior and posterior probabilities have been strongly questioned; see, for example, Stangl (1996). Many analysts actually interpret them as the (prior and posterior) probability that the corresponding model is the 'true' mechanism generating the phenomenon under study. However, according to Chatfield (1995) and references therein, the existence of a 'true' underlined model is rarely a realistic assumption and in fact 'no-one really believe this'.

Bernardo and Smith (1994) adopt three approaches for the interpretation of prior and posterior model probabilities:

1. $\mathcal{M}$-*closed* view : $m_T \in \mathcal{M}$; where m_T is the unknown 'true' underlined model.

2. $\mathcal{M}$-*completed* view: $\mathcal{M}$ is simply a set of specified models for comparison, 'to be evaluated in the light of the individuals separate actual belief model'.

3. $\mathcal{M}$-*open* view: $\mathcal{M}$ here is simply a set of specified models for comparison, with 'no separate overall actual belief specification'.

In $\mathcal{M}$-*closed* view the interpretation of $f(m|\boldsymbol{y})$ as the posterior probability that model m is the 'true' mechanism generating the phenomenon under study is coherent and valid but Bernardo and Smith (1994) argue that a real underlined model usually does not exist and therefore this view is not realistic unless in extreme cases such as evaluation of a 'computer game'. When $\mathcal{M}$-*closed* view cannot be adopted, our aim is to identify a good 'proxy' of the real model or simply consider which model performs best (in terms of prediction or data fitting) over the selected set of models. For this reason, the alternative term (prior and posterior) 'model weight' can be used instead of probability. A different interpretation may be adopted: (prior or posterior) model weights are the (prior or posterior) probabilities that the corresponding model is the 'best' approximation or description of reality over the selected set of candidate models $\mathcal{M}$.

2.2.1.5 Model Selection and Rejection as a Decision Problem

Within the Bayesian framework, model selection and hypothesis testing is viewed as a decision problem. Although posterior model probabilities (or weights) play an important role in the model selection or rejection procedure, we additionally need to specify utilities over which the final decision will be based. We may assign utilities $u'(m_T, m)$ and select the model which maximizes the posterior expected utility $E_{m|y}[u'(m_T, m)]$. The function $u'(m_T, m)$ denotes the utility when m_T is the true model but we select model m instead. Usual utilities in model selection are

- $u'(m, m) = 1$ and $u'(m, m') = 0$ for all $m' \neq m$, $m', m \in \mathcal{M}$.
- $u'(m_k, m_l) = -u'_{kl}$ where u'_{kl} are constant positive values.
- Kullback-Leibler discrepancy of the posterior distributions of two competing models. This discrepancy is a distance measure between two distributions defined as

$$K(f, g) = \int log \left[\frac{f(x)}{g(x)}\right] f(x)dx. \tag{2.6}$$

The first case leads to the use of posterior odds and to a Bayesian test which selects model m_0 if $PO_{01} > 1$ and m_1 otherwise. The second case facilitates again the posterior odds but with different cut-off point. Therefore we now select the null model when $PO_{01} > [u'_{01} - u'_{11}]/[u'_{10} - u'_{00}]$ for the sensible choices of $u'_{00}, u'_{11} < u'_{10}, u'_{01}$. Berger *et al.* (1994, 1997) constructed more sophisticated tests that have dual interpretation in both classical and Bayesian statistics and an additional area where decision can be taken (both models are equally good). For details see Berger *et al.* (1994, 1997).

Model selection based on more complicated utility functions (usually the Kullback-Leibler discrepancy) has been used by San Martini and Spezzaferri (1984), Poskitt (1987), Bernardo and Smith (1994), Key (1996), Key *et al.* (1997, 1998).

Recent work involves the Bayesian reference criterion (BRC) proposed by Bernardo (1999) derived by the use of Kullback-Leibler discrepancy as utility. For $\boldsymbol{\beta}^T = [\boldsymbol{\beta}'^T, \boldsymbol{\beta}''^T]$ and the assessment of $H_0 : \boldsymbol{\beta}' = \boldsymbol{\beta}'_0$ $vs.$ $H_1 : \boldsymbol{\beta}' \neq \boldsymbol{\beta}'_0$. The selection procedure for such case is given by three steps:

[1] Compute the Kullback-Leibler distance between the models defined in the two hypotheses: $K[f(\boldsymbol{y}|\boldsymbol{\beta}', \boldsymbol{\beta}''), f(\boldsymbol{y}|\boldsymbol{\beta}'_0, \boldsymbol{\beta}'')]$.

[2] Compute the posterior expectation of the above distance

[3] Reject H_0 when the posterior expectation computed in the second step is greater than a critical value d^*; the value $d^* = 5$ was proposed by Bernardo (1999) for scientific communication.

Goutis and Robert (1998) also used Kullback-Leibler distance in hypothesis testing. They proposed to test the hypothesis $H_0 : K(f(.|\boldsymbol{\theta}), f(.|\boldsymbol{\theta}_0)) \leq d^*$ *vs.* $H_1 : K(f(.|\boldsymbol{\theta}), f(.|\boldsymbol{\theta}_0)) > d^*$ and implement this approach in generalised linear models using MCMC.

2.2.2 Bayesian Model Averaging and Prediction

Bayesian theory offers the tool to adjust predictions (and inference) according to the observed model uncertainty. This methodology is called *Bayesian model averaging* since the distribution of any quantity of interest [for example $f(\boldsymbol{\beta}|\boldsymbol{y})$] is now the average of all conditional model specific posterior distributions [$f(\boldsymbol{\beta}|m, \boldsymbol{y})$] weighted by their posterior model probabilities $f(m|\boldsymbol{y})$. By this way, Bayesian methods base their predictions on all models under consideration and therefore account for model uncertainty. Similarly, the predictive distribution is given by

$$f(\Delta|\boldsymbol{y}) = \sum_{m \in \mathcal{M}} f(\Delta|m, \boldsymbol{y}) f(m|\boldsymbol{y}) \tag{2.7}$$

where Δ is a parameter of interest and $f(m|\boldsymbol{y})$ is given by formula (2.2).

Wasserman (1997) and Hoeting *et al.* (1998) recently provided two well written papers that both review Bayesian model averaging. According to Hoeting *et al.* (1998) the idea of model averaging seems to exist from the beginning of nineteenth century in the early work of Laplace but was firstly formulated by Leamer (1978) without gaining great attendance due to computation difficulties. After the reinvention of MCMC, Madigan and Raftery (1994) and Kass and Raftery (1995) brought again in the foreground Bayesian model averaging. General theoretical and practical details on Bayesian model averaging are given by Kass and Raftery (1995), Madigan *et al.* (1996), and Hoeting *et al.* (1998). Comprehensive discussion on model uncertainty and model averaging is also provided by Draper (1995) and Chatfield (1995). Implementation of Bayesian model averaging in linear models is given by Raftery *et al.* (1997), while Hoeting *et al.* (1995, 1996) provide Bayesian model averaging using simultaneous variable and outlier or transformation identification in linear models. Bayesian model averaging methods were also applied by York *et al.* (1995) in estimating proportion of born children with Down's syndrome, Clyde *et al.* (1996) in generalised linear models using an alternative orthogonal model space, Fernandez *et al.* (1997) in modelling fishing activities, Heckerman and Meek (1997) in Bayesian regression and classification models in networks, Clyde and DeSimone-Sasinowska (1997) in Poisson models, Clyde *et al.* (1998) in wavelets and Clyde (1999) in linear and generalised linear models using some clever MCMC samplers for approximating posterior weights. Finally, Fernandez *et al.* (1998) provide benchmark priors for Bayesian model averaging. Buckland *et al.* (1997) try to simplify Bayesian model averaging techniques by calculating the model weights through AIC and BIC approximations.

The predictive performance of any model is usually measured by the *logarithmic scoring rule* (LS) which is given by

$$LS = -E\left\{log\left[\sum_{m\in\mathcal{M}} f(\Delta|m,\boldsymbol{y})f(m|\boldsymbol{y})\right]\right\} \tag{2.8}$$

for Bayesian model averaging and by

$$LS_m = -E\{log[f(\Delta|m,\boldsymbol{y})]\} \tag{2.9}$$

for model m; Δ denotes a future observation and the expectation is with respect to $f(\Delta|\boldsymbol{y})$. Lower values of the logarithmic scoring rule indicate better predictive power. The Bayesian model averaging method always provides better predictive ability (in terms of logarithmic scoring rule) since $LS \leq LS_m$, $\forall\, m \in \mathcal{M}$; see Madigan and Raftery (1994), Kass and Raftery (1995) and Raftery *et al.* (1997). Implementation of assessing predictive performance using logarithmic scoring rule is provided by Madigan *et al.* (1995), Raftery *et al.* (1996), Volinsky *et al.* (1997) and Hoeting *et al.* (1998).

2.2.3 Occam's Window

In many cases averaging over all possible models is not possible for example when the number of models under consideration is large. An alternative is to average over a limited set of 'best' models. This variant of model averaging was called by Heckerman and Meek (1997) 'selective model averaging'. One method for identifying the most promising models is suggested by Madigan and Raftery (1994). This method is called *Occam's window* and is based on the Occam's razor logic, widely used in other disciplines; for implementation of Occam's razor in astronomy see Jefferys and Berger (1991). Although averaging over $\mathcal{A} \subset \mathcal{M}$ may give different predictions than original Bayesian model averaging, there is some evidence that it will still have better predictive power than the selection of a single model. In some cases Occam's window may ignore a wide range of uncertainty since it is usual to have many models each of them with low posterior weights but their union set accounts for important percentage of the total uncertainty.

Occam's window is not an MCMC method. It is a simple and fast algorithm that restricts our attention to a set of the most promising models. The algorithm sets two conditions. The first one ignores all models that are far away (in terms of posterior odds) from the best model. The second condition ignores models which are more complicated and worse in terms of posterior odds than at least one model selected by the first condition. In detail we firstly select

$$\mathcal{A}' = \left\{ m \in \mathcal{M} : \frac{max_{m_l}\{f(m_l|\boldsymbol{y})\}}{f(m|\boldsymbol{y})} \leq \kappa^* \right\} \tag{2.10}$$

where κ^* is a constant that according to Kass and Raftery (1995) should be equal to 20 (by analogy to the popular 0.05 cut-off for p-values). Then we exclude a set of models $\mathcal{B}$ given by

$$\mathcal{B}' = \left\{ m \in \mathcal{M} : \ \exists \ m_l \in \mathcal{A}', \ \mathcal{V}(m_l) \subset \mathcal{V}(m), \ \frac{f(m_l|\boldsymbol{y})}{f(m|\boldsymbol{y})} > 1 \right\} \tag{2.11}$$

that is, the set $\mathcal{B}'$ will include any model m for which there are simpler sub-models m_l with higher posterior probability. Note that one model m_l is sub-model of m if the former includes all terms of the latter. Finally, we set $\mathcal{A} = \mathcal{A}' \setminus \mathcal{B}'$ and calculate the predictive density

$$f(\Delta|\boldsymbol{y}) = \sum_{m \in \mathcal{A}} f(\Delta|m, \boldsymbol{y}) f(m|\boldsymbol{y})$$

instead of equation (2.7).

In situations that we have to compare two models m_0 and m_1 we may interpret the posterior odds of model m_0 against model m_1 (PO_{01}) as follows.

- If $PO_{01} > \kappa^*$ then reject m_0 and consider m_1. We can interpret this case as strong evidence in favour of m_0 and/or against m_1 .

- If $PO_{01} \in (1, \kappa^*)$ then we generally consider as equally 'good' both models since there is not strong enough evidence in favour of m_0. In the case where m_0 is sub-model of m_1 then we may consider only the simpler model m_0 due to the second condition of Occam's window.

- If $PO_{01} \in (\kappa^{*-1}, 1)$ then we consider both models as equally 'good'. There is evidence against m_0 but not strong enough in order to reject it.

- If $PO_{01} < \kappa^{*-1}$ then reject m_0 and consider m_1 since there is strong evidence against model m_0 and/or in favour of m_1.

Two search algorithms where provided by Madigan and Raftery (1994) for graphical models - the *Up* and *Down* algorithms. When we start from a non-saturated or a non-empty model then we execute firstly the Down and then the Up algorithm. Let $\mathcal{A}$ and $\mathcal{C}$ be subsets of the model space $\mathcal{M}$, where $\mathcal{A}$ denotes a set of 'promising models' and $\mathcal{C}$ the models under consideration in each step of the algorithm. For both algorithms we start with $\mathcal{A} = \emptyset$ and $\mathcal{C} = \mathcal{M}$.

Down Algorithm

1. Select model $m \in \mathcal{C}$.
2. $\mathcal{C} \leftarrow \mathcal{C} \setminus \{m\}$ and $\mathcal{A} \leftarrow \mathcal{A} \cup \{m\}$.
3. Select a sub-model m_0 of m by removing a variable/link from m.
4. Compute $LPO = log(PO_{m_0,m})$.
5. If $LPO > log(\kappa^*)$ then $\mathcal{A} \leftarrow \mathcal{A} \setminus \{m\}$ and if $m_0 \notin \mathcal{C}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{m_0\}$.
6. If $-log(\kappa^*) \le LPO \le log(\kappa^*)$ then if $m_0 \notin \mathcal{C}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{m_0\}$.
7. If there more submodels of m, go to 3.
8. If $\mathcal{C} \neq \emptyset$, go to 1.

Up Algorithm

1. Select model $m \in \mathcal{C}$.
2. $\mathcal{C} \leftarrow \mathcal{C} \setminus \{m\}$ and $\mathcal{A} \leftarrow \mathcal{A} \cup \{m\}$.
3. Select a supermodel m_1 of m by adding a variable/link to m.
4. Compute $LPO = PO_{m,m_1}$.

5. If $LPO < -log(\kappa^*)$ then $\mathcal{A} \leftarrow \mathcal{A} \setminus \{m\}$ and if $m_1 \notin \mathcal{C}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{m_1\}$.

6. If $-log(\kappa^*) \leq LPO \leq log(\kappa^*)$ then if $m_1 \notin \mathcal{C}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \{m_1\}$.

7. If there more supermodels of m, go to 3.

8. If $\mathcal{C} \neq \emptyset$, go to 1.

More details are given in Madigan and Raftery (1994) and Kass and Raftery (1995). Application in normal linear models is presented by Raftery *et al.* (1997) and in proportional hazard models by Volinsky *et al.* (1997). Raftery (1995) applied Occam's window in social sciences and Raftery and Richardson (1996) in epidemiology.

2.2.4 Lindley's Paradox

Lindley (1957) reported a strange phenomenon on the behaviour of posterior odds. He used the simple example where $y \sim N(\theta, \sigma^2)$ with σ^2 known, $H_0 : \theta = \theta_0$ *vs.* $H_1 : \theta \neq \theta_0$. He assigned prior probability $p = P(H_0)$ to H_0 and on $\theta|H_1$ a uniform over an interval I containing θ_0 and $\bar{y}$ 'is well within the interval I'. The resulting posterior odds is given by

$$PO_{01} = \frac{p}{1-p} \frac{exp(-\frac{n}{2\sigma^2}(\bar{y}-\theta_0)^2)}{\sqrt{2\pi}\sigma/\sqrt{n}}.$$

The resulting posterior odds depends on the sample size n and $\bar{y}$. Lindley considered samples being at the limit of rejection area of the usual significance test of $100q\%$ significance level. These samples have $\theta = \theta_0 \pm z_{q/2}\sigma^2/\sqrt{n}$ (or z_q for one tailed alternatives) and resulting posterior odds

$$POLS_{01}^q = \frac{p}{1-p} \frac{exp(-\frac{1}{2}z_{q/2}^2)}{\sqrt{2\pi}\sigma/\sqrt{n}}$$

where z_q is the q quantile of the standardised normal distribution. We will use the term 'posterior odds at the limit of significance'. From the above, he noted that when n increases the above posterior odds also increase and tend to infinity for a given significance level q. This leads to a paradox since for sufficient large samples Bayesian methods and significance tests support different hypotheses.

Bartlett (1957) noted another paradox which is related to the prior variance used in Bayes factor. He observed that the largest the prior variance is, the largest the posterior Bayes factor in favour of H_0 will be. This phenomenon is much of concern since usual 'improper' priors cannot be determined due to unknown constants involved in the computation of posterior odds while large variance priors fully support the simplest model. For these reasons either improper or large variance priors cannot be used. Although, Bartlett (1957) noted this phenomenon, the term 'Lindley's paradox' is used for any case where Bayesian and significance tests result in contradictive evidence (Shafer, 1982). The term 'Bartlett' paradox was used by a few number of researchers such as Kass and Raftery (1995) while others refer to this phenomenon by the term 'Jeffreys' paradox (Lindley, 1980, Berger and Delampady, 1987) or 'Jeffreys-Lindley's paradox' (Robert, 1993). Detailed discussion of Lindley's paradox is provided by Shafer (1982).

Although Lindley (1993) noted that the sensitivity of the Bayes factor is natural, this drawback resulted in a series of publications trying to resolve the problem and find good reference priors for model selection. In this category we may include Bayes factor variants (posterior, fractional and intrinsic) which are defined in the following section. Other approaches are presented by Robert (1993) and Brewer (1998); see also Chapter 6 of this thesis. Berger and Selke (1987), Casella and Berger (1987), Berger and Delampady (1987), Delampady and Berger (1990) and Berger and Montera (1991) examine the relationship of p-values and posterior probabilities using precise ($H_0 : \theta = \theta_0$) and imprecise ($H_0 : |\theta - \theta_0| \leq \epsilon$) hypotheses. They used the lower bounds of Bayes factors to make comparisons with classical statistics. Casella and Berger (1987) argue that the main reason for Bartlett's paradox is that we incoherently put mass point prior on a single point of the parameter space and this 'actually reflects a bias towards H_0'.

2.2.5 Bayes Factors Variants

The need for use of non-informative priors in model selection led to the definition of three new types of Bayes factors: the posterior, fractional and intrinsic Bayes factors by Aitkin (1991), O'Hagan (1995) and Berger and Pericchi (1996a, 1996b), respectively.

The posterior Bayes factor is given by

$$PBF_{01} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m_0)}, m_0) f(\boldsymbol{\beta}_{(m_0)}|m_0, \boldsymbol{y}) d\boldsymbol{\beta}_{(m_0)}}{\int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m_1)}, m_1) f(\boldsymbol{\beta}_{(m_1)}|m_1, \boldsymbol{y}) d\boldsymbol{\beta}_{(m_1)}}$$

which essentially is the mean likelihood under the given posterior while the original Bayes factor is the mean likelihood over the selected prior. Posterior Bayes factor allows the use of improper priors but is incoherent since it uses the data information twice and therefore violates the likelihood principle. Moreover, it is not derived from the Bayes theorem and therefore it cannot be considered as a 'pure' Bayesian tool.

The fractional Bayes factor was introduced by O'Hagan and is given by

$$FBF_{b,01} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m_0)}, m_0)^{1-b} f_b(\boldsymbol{\beta}_{(m_0)}|m_0, \boldsymbol{y}) d\boldsymbol{\beta}_{(m_0)}}{\int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m_1)}, m_1)^{1-b} f_b(\boldsymbol{\beta}_{(m_1)}|m_1, \boldsymbol{y}) d\boldsymbol{\beta}_{(m_1)}}$$

where

$$f_b(\boldsymbol{\beta}_{(m)}|m_0, \boldsymbol{y}) d\boldsymbol{\beta}_{(m)} = \int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m)^b f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y}) d\boldsymbol{\beta}_{(m)}$$

and $b < 1$ is called fractional parameter. The idea of fractional Bayes factor is based on the notion of using a fraction of the likelihood for estimation and the rest for model selection. Although, the fractional Bayes factor is a useful alternative statistical tool for model selection and in some cases may provide valuable insight, is also not 'purely' Bayesian.

The intrinsic Bayes factor of Berger and Pericchi (1996a,b) was based on the original idea of Spiegelhalter and Smith (1982) of partial Bayes factor in which we use a small fraction of the data for estimation and the rest of the data for model selection. Given a training sample $\boldsymbol{y}(l)$ the partial Bayes factor is given

$$B_{01}(\boldsymbol{y}(l)) = \frac{\int f(\boldsymbol{y}(\backslash l)|\boldsymbol{\beta}_{(m_0)}, m_0) f(\boldsymbol{\beta}_{(m_0)}|m_0, \boldsymbol{y}(l)) d\boldsymbol{\beta}_{(m_0)}}{\int f(\boldsymbol{y}(\backslash l)|\boldsymbol{\beta}_{(m_1)}, m_1) f(\boldsymbol{\beta}_{(m_1)}|m_1, \boldsymbol{y}(l)) d\boldsymbol{\beta}_{(m_1)}}$$

where $\boldsymbol{y}(\backslash l)$ is the rest of data used for model selection. The intrinsic Bayes factor is estimated by the median, arithmetic or geometric mean of partial Bayes factors over all minimal samples.

Implementation of the posterior Bayes factor is provided in distribution fitting using the exponential distribution family (Aitkin, 1995) and in selection of normal mixture distributions (Aitkin *et al.* , 1996). A comparison between p-values, posterior Bayes factors and AIC criterion is provided by Aitkin (1997). Further work on the fractional Bayes factor is provided by Conigliani and O'Hagan (1996) and DeSantis and Spezzaferri (1997) while intrinsic Bayes factor has been applied to autoregressive data by Varshavsky (1996). Properties of both intrinsic and fractional Bayes factors are given by O'Hagan (1997). Berger and Pericchi (1999) propose the median version of intrinsic Bayes factor as a well behaved measure. Moreover, Berger and Pericchi (1998) compare common Bayes factor and its approximations with its intrinsic and fractional variants. Finally, Berger and Montera (1998) use intrinsic and fractional Bayes factors with one sided alternative hypotheses while Berger and Pericchi (1998) compare both intrinsic and fractional Bayes factor with the usual (prior) Bayes factor and its approximations.

2.2.6 Bayesian Predictive Model Selection

Alternative methods for model assessment and model adequacy, rather than model selection, are the predictive measures. In this category we may include the criteria of Ibrahim and Laud (1994) and Bayesian variants of p-values (Dempster, 1974, Box, 1980, Rubin, 1984, Bayarri and Berger, 1998a,b,c).

2.2.6.1 Predictive Model Selection Criteria

Predictive model selection (or rejection) use the predictive distributions of type $f(y_{n+1}|\boldsymbol{y}, m)$ in order assess whether model m describes sufficiently the data; y_{n+1} here denote future observations. Many predictive criteria have been developed, see, for example, Bernardo and Smith (1984), Key (1996) and references therein. For illustrative purposes we will briefly concentrate on a recent approach introduced by Ibrahim and Laud (1994) for illustration.

Ibrahim and Laud (1994) propose to use predictive distributions for variable selection and model assessment. Suppose that we replicate the entire experiment with the same design matrix $\boldsymbol{X}$. The predictive distribution of the vector of responses that we might obtain from this replicated experiment is given by

$$f(\boldsymbol{z}|m, \boldsymbol{y}) = \int f(\boldsymbol{z}|\boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y}) d\boldsymbol{\beta}_{(m)}. \tag{2.12}$$

A good measure of the model fit is given by L_m criterion defined as

$$L_m^2 = E_{\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}_{(m)}}[(\boldsymbol{z} - \boldsymbol{y})^T(\boldsymbol{z} - \boldsymbol{y})] \tag{2.13}$$

which measures the expected squared distance of the replicated and the observed data. The smallest values of L_m indicate better models. This L_m criterion is measured in the same units as the response variable Y.

Another criterion used by Laud and Ibrahim (1995) is given by

$$M_m^* = f(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{X}_{(m)})$$

which is the probability of getting again the same replicated data. This is equivalent to posterior Bayes factor as defined by Aitkin (1991). The $M_m^{*-1/n}$ is measured in the same units as the response variable Y; where $M_m^{*-1/n}$ is the M^* in the power of $-1/n$.

Finally, the $K_m^{m_0}$ criterion is defined as the sum of two Kullback-Leibler discrepancies and is given by

$$K_m^{m_0} = K[f(\boldsymbol{z}|m_0, \boldsymbol{y}), f(\boldsymbol{z}|m, \boldsymbol{y})] + K[f(\boldsymbol{z}|m, \boldsymbol{y}), f(\boldsymbol{z}|m_0, \boldsymbol{y})]$$

where m_0 is a fixed model defined for comparison. Usually m_0 is either the constant or the full model. The quantity $K[f(\boldsymbol{z}|m_1, \boldsymbol{y}), f(\boldsymbol{z}|m_2, \boldsymbol{y})]$ is the Kullback Leibler distance between the predictive densities of the two models as defined in (2.6). Both these L_m and M_m^* criteria measure how close is the predictive density of model m to the observed data while the third one ($K_m^{m_0}$) measures how close is the predictive density of model m to the corresponding predictive density of model m_0.

Implementation of this predictive selection approach has been provided in designed experiments (Ibrahim and Laud, 1994), linear regression models and transformation selection (Laud and Ibrahim, 1995, Ibrahim and Laud, 1996, Hoeting and Ibrahim, 1997), multivariate linear model (Ibrahim and Chen, 1997) and repeated measures random effects models (Weiss *et al.* , 1997).

2.2.6.2 Bayesian Predictive P-Values

Other predictive measures are the 'Bayesian' p-values which are similar in notion to the classical p-values. Generally, p-values are measures of surprise in the data relative to the the hypothesized model (Bayarri and Berger, 1998a,b,c). The p-values have the following form

$$p = P(T(Y) > T(\boldsymbol{y})|m),$$

where $T(Y)$ is a test statistic, $T(\boldsymbol{y})$ is the observed value of this statistic and m is the hypothesized model. Variation of p-values depend on the distribution used for the calculation of this tail area probability. The most commonly used p-values are

1. *Classical p-values* in which we use the maximum likelihood $f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m)}, m)$.

2. *Prior predictive p-values* (Box, 1980) in which we use the prior predictive density defined by
$$f(\boldsymbol{z}|m) = \int f(\boldsymbol{z}|\boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|m) d\boldsymbol{\beta}_{(m)}.$$

3. *Posterior predictive p-values* (Guttman, 1967, Rubin, 1984, Meng, 1994) in which we use the posterior predictive density defined by (2.12).

4. *Conditional predictive p-values* (Bayarri and Berger, 1998a,b,c) which is a compromise between the two previous p-values. This p-value is calculated in respect to
$$f(T(Y)|U(\boldsymbol{y}), m) = \int f(T(Y)|U(\boldsymbol{y}), \boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|U(\boldsymbol{y}), m) d\boldsymbol{\beta}_{(m)},$$
where $U(\boldsymbol{y})$ is the observed value of a second test statistic $U(Y)$.

5. *Partial Posterior P-values* (Bayarri and Berger, 1998b) which is calculated in respect to the *partial posterior*
$$f(T(Y)|U(\boldsymbol{y}), m) \propto \int f(T(Y)|\boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{y}|T(\boldsymbol{y}), \boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|m) d\boldsymbol{\beta}_{(m)}.$$

All p-values are strongly criticized and do not provide at any case the probability that the hypothesized model is true. According to Meng (1994) p-values are only measures of discrepancy between data and the model, similar to L_m and M^*_m predictive criteria defined in the previous section. The same author also adds that p-values may be useful 'in monitoring model adequacy'. Similar ideas are expressed by Lewis and Raftery (1996) supporting that 'posterior predictive assessment should usually be used to point the way to a better model, rather than to reject the current model in the absolute sense'. Moreover, Bayarri and Berger (1998c) claim that p-values are important in deciding whether to search for alternative model but their use as measures of model rejection is questionable.

The main disadvantages of the prior predictive p-values are their sensitivity to the prior distribution and that cannot be specified when improper priors are used. The main drawback of the posterior p-values is that they are not 'Bayesian' since they use the data twice and violate the likelihood principle (Meng, 1994). This double use of the data leads to overestimated measures that are high even when the test statistic is far away from the hypothesized model (see Bayarri and Berger, 1998c, for illustration). Also for suitably large n and non-informative priors the posterior p-values give equivalent results in classical p-values.

This was considered as a disadvantage from Bayarri and Berger (1998c) but seems to be in concordance with the use of non-informative priors in Bayesian estimation problems that result in posterior distributions equivalent to the maximum likelihood estimates. On the other hand, they are not sensitive on the choice of prior distribution and improper priors can be used. Bayarri and Berger (1998c) developed alternative conditional predictive p-values that facilitate the advantages and avoid the drawbacks of both prior and posterior p-values. The only drawback is that we need to select a second statistic upon which we need to condition on. The selection of this statistic may alter the results but the author provide adequate guidance for its selection. They also provide various alternative computational schemes for the estimation of these p-values including Gibbs and Metropolis-Hastings sampler. The partial posterior predictive p-value have properties similar to the conditional posterior p-values and furthermore they do not use any arbitrary statistic upon we need to condition on; for further discussion see Bayarri and Berger (1998b).

Other publications include the p-values by Dempster (1974) and Aitkin (1997) discussing posterior predictive values based on the statistic of the likelihood ratio test. Aitkin (1997) also provides connection of posterior p-values using the maximum likelihood ratio with the classical p-values, posterior Bayes factor and information criteria such as AIC. Gelman *et al.* (1996) and Gelman and Meng (1996) provide algorithms for the calculation of posterior p-values and other predictive discrepancies using MCMC output. Similarly, Spiegelhalter *et al.* (1996a) provide guidance for the estimation of posterior p-values and other predictive measures using BUGS.

Other comprehensive measures of surprise are given by Evans (1997) and Bayarri and Berger (1998a).

2.3 Model Selection Criteria

The use of information criteria in model choice was introduced in the early seventies in order to find a consistent method for model selection. The most popular criteria are Akaike's Information Criterion (AIC, Akaike, 1973), Bayes Information Criterion (BIC, Schwarz, 1978) and C_p Criterion (Mallows, 1973). These criteria have been widely used by 'frequentists' although most of them such as BIC (Schwarz, 1978) derive from Bayesian logic. BIC was derived as a 'large sample approximation' of Bayes factor while AIC as an approximately unbiased estimate of the Kullback-Leibler discrepancy between two models (Akaike, 1973). These criteria have been mainly used for linear models and a selective variety is presented in Table 2.3. In this table, RSS_m is the maximum likelihood residual sum of squares for model m equal to

$$RSS_m = \left(\boldsymbol{y} - \boldsymbol{X}_{(m)}\hat{\beta}_{(m)}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)}\right),$$

where $\hat{\boldsymbol{\beta}}_{(m)}$ are the maximum likelihood parameter estimates of model m, $d(m)$ is the dimension of the parameter vector $\boldsymbol{\beta}_{(m)}$ and $\hat{\sigma}^2$ is the maximum likelihood estimate of the residual variance.

Generally, most information criteria minimize the quantity

$$IC_m = -2log\left(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_m, m)\right) + d(m)F \tag{2.14}$$

where $\boldsymbol{\theta}_m$ is the whole parameter vector, $\hat{\boldsymbol{\theta}}_m$ are the corresponding maximum likelihood estimates. In linear regression models $\boldsymbol{\theta}_m^T = [\boldsymbol{\beta}_{(m)}^T, \sigma^2]$ and minimizing $-2log\left(f(\boldsymbol{y}|\hat{\theta}_m, m)\right)$ is equivalent to minimizing $nlog(RSS_m)$. Also note that F is the penalty imposed to the -2log-likelihood for each additional parameter used in the model. Different penalty functions result in different criteria; for example

- For $F = 2$ we have AIC.
- For $F = 3/2$ we have the criterion of Smith and Spiegelhalter (1980).
- For $F = log(n)$ we have BIC.
- For $F = clog(log(n))$ we have Φ_c.

If we want to compare two models m_0 and m_1 then we select the one that has lower value of IC and therefore we define as IC_{01} the difference of the two information criteria. Hence

$$IC_{01} = -2log\left(\frac{f(\boldsymbol{y}|\hat{\theta}_{m_0}, m_0)}{f(\boldsymbol{y}|\hat{\theta}_{m_1}, m_1)}\right) - [d(m_1) - d(m_0)]F. \tag{2.15}$$

Without loss of generality, we assume that $d(m_0) < d(m_1)$. Note that if $IC_{01} < 0$ we select model m_0 and if $IC_{01} > 0$ we select model m_1. We can generalise the above criterion difference by substituting the expression $[d(m_1) - d(m_0)]F$ by a more complicated penalty function ψ. In such case we may write the information criteria in more general setup given by

$$IC_{01} = -2log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_0}, m_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_1}, m_1)}\right) - \psi,$$

where ψ is a penalty function depending on difference of model dimensionalities, $d(m_1) - d(m_0)$, sample size n, and design matrices, $\boldsymbol{X}_{(m_0)}$ and $\boldsymbol{X}_{(m_1)}$.

Shao (1997) divides model choice criteria in three major divisions:

1. Asymptotically valid criteria under the assumption that a true model exists.
2. Asymptotically valid criteria under the assumption that not a fixed dimension true model exists.
3. A compromise between these two categories.

According to Zhang (1997) the main conclusion of Shao (1997) is that IC_m with $F = 2$ and $F \to \infty$ when $n \to \infty$ are two differently behaved categories of criteria referred as AIC-like and BIC-like criteria. The BIC-like criteria perform better if the true model has simple structure ('finite dimension') while the AIC-like criteria are better if the true model is a more complex one ('infinite dimension'). The main argument of Zhang (1997) in favour of BIC-like criteria is that the existence of a true model is doubtful and even if exists we may prefer to select a simpler model that approximates sufficiently the true one. In his words, 'the practical advantage of a more parsimonious model often overshadows concerns over the correctness of the model. After all the goal of statistical analysis is to extract information rather to identify the correct model.' In this direction, Rissanen (1986) states that it is obvious that all selection criteria give rise to quantification of the parsimony principle. They differ in the weight (or significance) that they give to goodness of fit and model complexity. The goodness of fit is measured by the log-likelihood ratio while model complexity by the number of model parameters. Zheng and Loh (1995) examine a generalization of information criteria as given in (2.14) and prove that BIC-like criteria are optimal and consistent. Bhansali (1997) supports Shao (1997) and Zhang (1997) claiming that when the dimension of the true model is 'finite' then AIC and final prediction error (FPE) do not provide consistent estimates of the true model while BIC and the criterion of Hannan and Quinn (1979) do provide consistent estimates. Moreover, in the case of 'infinite' dimension of the true model

AIC, final prediction error and Shibata criterion (1980, 1981) are asymptotically efficient while BIC is not.

Bhansali and Downham (1977) introduced an AIC variant given by (2.14) with $2 \le F \le 5$. They argued that optimal values of the penalty may be greater than 2 using frequency tabulations of generated data. According to Akaike (1979) and Atkinson (1980) the use of frequency tabulations to prove optimality of a criterion is not appropriate. Akaike (1979) alternatively proposed to use squared prediction error. Shi and Tsai (1998) used Kullback-Leibler discrepancy to define other variants of AIC. Also Akaike (1981) facilitates Bayesian theory to result in alternative criteria.

Wei (1992) proposes another criterion called Fisher information criterion (FIC) where we minimize

$$FIC_m = \sigma^2_m (n + log|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}|).$$

If we substitute σ^2 by its maximum likelihood estimate, then minimizing the above quantity is equivalent to minimizing

$$nlog(FIC_m) = nlog(RSS_m) + log|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}|$$

when $n^{-1}log|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}| \rightarrow 0$. This criterion, which is equivalent to (2.14) with $F = d(m)^{-1}log|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}|$, was introduced for the normal model and can be adopted for other generalised linear models using the respective Fisher information matrix. Dudley and Haughton (1997) introduce information criteria based on Bayes factor with Jeffreys prior suitable for model selection when using multiple datasets. Lai and Lee (1997) use the final prediction error criterion which minimizes

$$FPE_m = \left(1 + \frac{2d(m)}{n}\right)\hat{\sigma}^2_m$$

which is equivalent to minimizing

$$nlog(FPE_m) = nlog(\hat{\sigma}^2_m) + 2d(m)$$

when $n \rightarrow \infty$ and therefore $2d(m)/n \rightarrow 0$. The above is the same as AIC. Similarly, minimizing

$$APE_m = n\hat{\sigma}^2_m + \hat{\sigma}^2 d(m) log(n)$$

is equivalent to using BIC. Shibata (1984) and Zheng and Loh (1997) extensively discuss the use of final prediction error as a model selection criterion. The latter also extents the final prediction error by substituting the dimensionality $d(m)$ by a function of it. Similarly, Zhang (1997) proposes an alternative criterion noted as $GIC^*_m = SS_m + F\hat{\sigma}^2 h[d(m)]$ which uses penalty for each unit of a dimensionality function; see, for example, Foster and George (1994) in which $h[d(m)] = log[d(m)]$. Stone (1977) and Shao (1993) also note that 'leave-one-out' cross-validation method is asymptotically equivalent to AIC and C_p.

Ronchetti (1997) examines the behaviour of a robust version of C_p and AIC criteria. The robust C_p (RC_p) 'allows us to choose the best models which fit *the majority of the data* by taking into account the presence of outliers and possible departures from the normality assumption on the error distribution'. Rao and Wu (1989) also propose a criterion similar to C_p.

Shibata (1997) gives a sufficient explanation why AIC was firstly proposed as a naive estimator of Kullback-Leibler discrepancy. He also discusses the bias correction given by Hurvich and Tsai (1989, 1991) which is more precise than Akaike's. Finally, he introduces bootstrap estimates of the Kullback-Leibler information. Cavanaugh and Shumway (1997) also introduced a bootstrap variant of AIC called WIC.

Geiger *et al.* (1996) extend BIC in Bayesian networks with hidden variables, where the dimension of the model is given by the rank of the Jacobian of the transformation between model parameters of the network and the parameters of observable variables.

Comparison of information criteria, posterior odds and likelihood ratios and their connections are provided by Atkinson (1981) and Chow (1981). Note that Bayes factor variants and Bayesian predictive or utility based criteria are (in most cases) equivalent to information criteria. For example the predictive criterion L_m of Ibrahim and Laud (1994) given by (2.13) is equivalent to an information criterion given by (2.15) with penalty function

$$F = \frac{n}{d(m_1) - d(m_0)} log\left(\frac{n - d(m_0) - 2}{n - d(m_1) - 2}\right)$$

while posterior Bayes factor is approximately equal to a criterion of type (2.15) with penalty equal to $F = log(2)$; see Aitkin (1991), O'Hagan (1995) and Chapter 6. Other recent interesting criteria are also presented by George and Foster (1997) using empirical Bayes methods, Bernardo (1999) using Bayesian decision theory and Gelfand and Ghosh (1998) using predictive loss approach.

	Full Name (Reference)	Equation
FPE	Final Prediction Error (Akaike, 1969)	$\left(1+\frac{2d(m)}{n}\right)\hat{\sigma}^2_m$
AIC	Akaike Information Criterion (Akaike, 1973)	(2.14), $F=2$
C_p	Mallows C_p (Mallows, 1973)	$\sigma^{-2}RSS_m - n + 2d(m)$
AIC$_a$	Generalization of AIC (Bhansali and Downham, 1977)	(2.14), $F=a, 2 \leq a \leq 5$
BIC	Bayes Information Criterion (Schwarz, 1978)	(2.14), $F=log(n)$
$\Phi_{c,m}$	Hannan and Quinn (1979) Criterion	(2.14), $F=clog(log(n))$
SSC	Smith and Spiegelhalter (1980) Criterion	(2.14), $F=3/2$
SC	Shibata (1980) Criterion	(2.14),$F=nlog[n+2d(m)]/d(m)$
BIVAR	Bias-Variance Criterion (Young, 1982)	$wC_p+(1-w)d(m)$
PMC	Predictive model Criterion	$BIC_{m_1}-BIC_{m_2}-2\log\bar{c}$
	(San Martini and Spezzaferri, 1984)	$\bar{c}=\frac{2(RSS_{m_1}/RSS_{m_2}-1)}{\log(RSS_{m_1}/RSS_{m_2})}-1$
FIC	Fisher Information Criterion (Wei, 1992)	$n\hat{\sigma}^2_m+\hat{\sigma}^2 log\lvert\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\rvert$
RIC	Risk Inflation Criterion (Foster and George, 1994)	(2.14), $F=2plog[d(m)]/d(m)$
EBC	Empirical Bayesian Criterion	$RSS_m/\hat{\sigma}^2+2d(m)logd(m)$
	(George and Foster, 1997)	$+2[p-d(m)]log[p-d(m)]$
		$-d(m)\{1+log[d(m)SS_m/\hat{\sigma}^2]\}$
BRC	Bayesian Reference Criterion (Bernardo, 1999)	see page 38

Table 2.3: Summary of Model Selection Criteria.

2.4 Discussion

In this chapter we have critically reviewed classical and Bayesian model selection methods. The main classical competitors are stepwise procedures using significance tests and evaluation of models via information criteria (mainly BIC and AIC). Alternatively, posterior odds (and Bayes factors) are the main tool for model selection and hypothesis testing in Bayesian analysis. Problems with prior specification together with the large computational burden, required for implementing Bayesian theory, did not allow posterior odds to become very popular until the early nineties. The development of MCMC methods and their implementation in statistical science made the calculation of posterior probabilities fairly automatic. Many authors (for example Bayarri and Berger, 1998a,b,c) have extensively studied relations between different model selection techniques. Current research mainly involves assessment of better criteria, evaluation of optimal MCMC methods and benchmark priors for model selection. In the following chapters we will introduce a new efficient, flexible and easy-to-use MCMC method for model selection; we will compare and underline relations between MCMC model selection methods; we will provide guidance for implementation of these methods in generalised linear models; finally we will describe a new perspective and interpretation of posterior odds.

Chapter 3

Model Selection via Markov Chain Monte Carlo Methods

3.1 Introduction

The problems with Bayesian model selection are associated with the computation of the integrals involved in the calculation posterior probabilities. These integrals can be analytically evaluated only in certain restricted examples. A further problem is that the size of the set of possible models $\mathcal{M}$ may be so great that calculation or approximation of $f(\boldsymbol{y}|m)$ for all $m \in \mathcal{M}$ becomes infeasible. Hence, MCMC methods for generating from the joint posterior distribution of $\left(m, \boldsymbol{\beta}_{(m)}\right)$ become an extremely attractive alternative. If a sample $(m^{(t')}, \boldsymbol{\beta}^{(t')},\ t' = 1, \ldots, t)$ can be generated from this distribution, then posterior model probabilities can be estimated directly by

$$\hat{f}(m) \;=\; \frac{1}{t}\sum_{t'=1}^{t} I(m^{(t')} = m) \qquad m \in \mathcal{M} \tag{3.1}$$

where $I(\cdot)$ is the indicator function. Samples from $f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})$ are also automatically available for marginal parametric inference. In practice, all suggested methods for generating from $f(m, \boldsymbol{\beta}_{(m)}|\boldsymbol{y})$ are based on Markov chains.

Furthermore, in many statistical models we may substitute the model indicator $m \in \mathcal{M}$ by $(s, \boldsymbol{\gamma}) \in \mathcal{S} \times \{0,1\}^p$, where the indicator vector $\boldsymbol{\gamma}$ represents which of the p possible sets of covariates are present in the model and s represents other structural properties of the model. For example, in generalised linear models, s may describe the distribution, link function and variance function, and the linear predictor may be written as

$$\boldsymbol{\eta} = \sum_{j \in \mathcal{V}} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j \tag{3.2}$$

where $\mathcal{V}$ is the set of possible regressors, $\boldsymbol{X}_j$ is the design matrix and $\boldsymbol{\beta}_j$ the parameter vector related to the jth term. When variable selection problems are considered then we may use either the model indicator m or the variable indicator vector $\boldsymbol{\gamma}$.

The latent vector of binary indicators $\boldsymbol{\gamma}$ was introduced by George and McCulloch (1993) in the first attempt to use MCMC algorithms for model selection. In their Gibbs based sampler, called 'stochastic search variable selection' (SSVS), they used the clever idea of keeping the dimensionality constant across all possible models by limiting the posterior distribution of non-significant (removed) terms in a small neighbourhood of zero instead of setting it equal to zero as usually. SSVS was originally implemented in linear regression models and was followed by a series of publications with implementation in various scientific fields; pharmacokinetics modelling (Wakefield and Bennett, 1996), construction of stock portfolios in finance (George and McCulloch, 1996), generalised linear models (George *et al.* , 1996, George and McCulloch, 1997), designed experiments (Chipman, 1996, 1997 and Chipman *et al.* , 1997) and multivariate regression models (Brown *et al.* , 1998).

Another popular method is the *Markov chain Monte Carlo model composition* (MC^3) which was originally implemented by Madigan and York (1995) in graphical model selection. Details of this MCMC algorithm is also given by Kass and Raftery (1995) while Ntzoufras (1995) provided extensive implementational details for log-linear model selection. MC^3 has been extended in other model selection problems such as variable, outlier and transformation identification in linear regression by Hoeting *et al.* (1995, 1996) and Raftery *et al.* (1997), social research by Raftery (1995), epidemiology by Raftery and Richardson (1996) and proportional hazards models by Volinsky *et al.* (1997).

These early research attempts were followed by Carlin and Chib (1995) sampler. Its computational complexity was its main drawback and the main reason for not being widely implemented in statistical research. An example of comparison between two models using Carlin and Chib method using BUGS software is given by Spiegelhalter *et al.* (1996c).

Green (1995) introduced a generalization of Metropolis-Hasting algorithm for sampling from models with different dimensionality, called *reversible jump* (RJ). This method is extremely flexible and can jump from one model space to another provided that we carefully select appropriate proposal densities. This sampler has gained attention and has been applied by Richardson and Green (1997) in selection of normal mixtures, Nobile and Green (1997) in ANOVA models and factorial experiments, Troughton and Godsill (1997) in autoregressive models, Dellaportas *et al.* (1998) in analysis of finite Poisson mixtures, Vrontos *et al.* (1998) in ARCH/GARCH models, Denison *et al.* (1998a) in CART models, Denison *et al.* (1998b) in MARS models, Giudici and Green (1998) in decomposable graphical Gaussian models and Dellaportas and Forster (1999) in log-linear models. Brooks and Giudici (1998) have developed convergence diagnostics for reversible jump output for samples within each model.

Other samplers are provided by Smith and Kohn (1996) for linear and nonparametric regression models, Clyde *et al.* (1996), Clyde (1999), Geweke (1996) and Kuo and Mallick (1998).

In this chapter, we focus on the prior specification for model selection used in generalised linear models and on methods of Green ('reversible jump', 1995) and Carlin and Chib (1995). We further consider existent 'variable selection' problems including SSVS and Kuo and Mallick samplers. Finally, we describe and comment on the fast variable selection methods used in normal models under a conjugate prior distribution.

3.2 Prior Specification

To complete Bayesian formulation of any model selection problem we define priors on parameters $\boldsymbol{\beta}_{(m)}$ and on model indicator m (or alternatively on $(s, \boldsymbol{\gamma})$). Here we review the most usual prior setups for generalised linear models. We focus on generalised linear models to facilitate the reading of the following chapters which involve applications of model selection methods on this popular family of models.

The usual prior specification in model selection is to define each model parameter prior distribution conditionally on the model indicator m. In variable selection we may use a prior on the parameter vector of the full model but, in most cases, this is not appropriate since the parameters have different interpretation under different models. This subsection is divided into three parts, the first one concerning priors for model parameters conditionally on the model indicator, the second concerning priors on model space while the last one introduces a different approach of prior distributions specification.

3.2.1 Prior Distribution for Model Parameters

A common prior distribution used to express prior beliefs about the model parameters is a multivariate normal distribution. Hence the prior is of the type

$$\boldsymbol{\beta}_{(m)}|m \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, \boldsymbol{\Sigma}_{(m)}\right), \tag{3.3}$$

where $\boldsymbol{\mu}_{\beta_{(m)}}$ and $\boldsymbol{\Sigma}_{(m)}$ are the prior mean and covariance matrix under model m respectively. This prior set up is used throughout this section. Due to the Lindley's and Bartlett's paradox described in Section 2.2.4 when no prior information is available, we should select a prior distribution with little information about model parameters $\boldsymbol{\beta}_{(m)}$ which is not extremely flat. The prior covariance matrix may be written alternatively as $\boldsymbol{\Sigma}_{(m)} = c^2\boldsymbol{V}_{(m)}$, where c^2 controls the flatness of the prior distribution and $\boldsymbol{V}_{(m)}$ encapsulates the prior correlation structure. Usual choice for the prior mean, when no prior information is available, is the zero vector, that is $\boldsymbol{\mu}_{\beta_{(m)}} = \mathbf{0}_{d(m)}$.

3.2.1.1 Independent Priors for Each Term Parameter Vector

In variable selection problems it is common to set independent priors on parameters of each variable or term. In such case the prior is given by

$$\boldsymbol{\beta}_j \sim N\left(\boldsymbol{\mu}_{\beta_j}, \boldsymbol{\Sigma}_j\right), \tag{3.4}$$

where $\boldsymbol{\mu}_{\beta_j}$ and $\boldsymbol{\Sigma}_j$ are the prior mean and covariance matrix for term j independent of the model indicator m (or $\boldsymbol{\gamma}$). Similar to the above prior setup, the covariance matrix can be alternatively written as $\boldsymbol{\Sigma}_j = c_j^2\boldsymbol{V}_j$ and for non-informative cases the prior mean is usually given by $\boldsymbol{\mu}_{\beta_j} = \mathbf{0}_{d_j}$. Geweke (1996) used independent truncated normal prior distributions in linear regression models.

This prior is plausible only when the design or data matrix is orthogonal since, in such cases, model parameters have similar interpretation over all models. We can easily incorporate such priors in ANOVA type models with sum-to-zero constraints. In other cases, when we are interested in prediction rather than description of variable relations, we may orthogonalize the design matrix and proceed with model selection in the new orthogonal model space; see Clyde *et al.* (1996). In non-orthogonal cases, especially when high dependencies among covariates exist, the use of such prior setup may result in unpredicted influence on the posterior odds and hence should be avoided. For a simple illustration see Chapter 6.

In the category of independent priors for term parameters we could include the prior of Dellaportas and Forster (1999) defined for log-linear model selection problems in contingency tables with sum-to-zero constraints. This prior is also proposed in this thesis for binomial logistic regression model selection problems with categorical data and can be extended in 'equivalent' priors for binomial models with other link functions through approximation of the Taylor expansion. The prior covariance structure of j term is given by

$$\boldsymbol{V}_j = d^{-1} \prod_{\nu \in \mathcal{V}(j)} (d_\nu + 1) \bigotimes_{\nu \in \mathcal{V}(j)} \left(\mathbf{I}_{d_\nu} - \frac{1}{d_\nu + 1} \mathbf{J}_{d_\nu} \right) \tag{3.5}$$

where $\mathcal{V}(j)$ is the set of factors creating the (interaction) term j, ν is a factor included in term j, d is the dimension of the full model (and the total number of cells in the contingency table), d_ν is the dimension of ν factor (and therefore $d_\nu + 1$ are the levels of this factor), $\mathbf{I}_{d_\nu}$ is the $d_\nu \times d_\nu$ identity matrix and $\mathbf{J}_{d_\nu}$ is the $d_\nu \times d_\nu$ matrix with every element equal to one. Dellaportas and Forster (1999) propose a value of $c_j^2 = c^2 = 2d$ (twice the number of cells) for Poisson log-linear models. Albert (1996) used similar prior with $\boldsymbol{V}_j = \boldsymbol{I}_{d_j}$ utilizing odds ratios to calibrate the prior parameter c. Therefore, his prior represents the statement that the corresponding odds ratio varies between $-2c$ and $2c$ with probability 0.68 when the two variables are not independent.

An equivalent prior for logistic regression problems with only categorical regressors can easily be adopted using Dellaportas and Forster (1999) prior. We simply consider the above prior covariance matrix multiplied by four since, in sum-to-zero constraints, the logistic regression parameters are twice the corresponding log-linear parameters. For all the other (non-canonical) link functions Taylor expansion may be used to find an equivalent prior variance; see Chapter 5.

3.2.1.2 Model Dependent Prior Distributions

In normal linear models the prior (3.3) can be used. Additionally we need to define a prior on the residual variance. The usual gamma distribution on the precision parameter $\tau = \sigma^{-2}$

(or inverse gamma on σ^2) may be considered. Therefore, the prior is given by

$$\tau \sim G(a_\tau, b_\tau). \tag{3.6}$$

An improper prior on the residual precision τ does not influence the posterior odds and hence (3.6) with $a_\tau = b_\tau = 0$ may be used without any complication. In normal linear models, it is convenient to use prior on model parameters conditionally on the residual variance σ^2. So, instead of (3.3) we may use

$$\boldsymbol{\beta}_{(m)}|\sigma^2, m \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, \boldsymbol{\Sigma}_{(m)}\sigma^2\right). \tag{3.7}$$

The joint prior distribution $f(\boldsymbol{\beta}_{(m)}, \sigma^2|m)$ given by the product of the marginal distributions (3.6) and (3.7) is also called normal inverse gamma distribution and it is conjugate since the posterior $f(\boldsymbol{\beta}_{(m)}, \sigma^2|m, \boldsymbol{y})$ is also normal inverse gamma; see, for example, Bernardo and Smith (1994) or O'Hagan (1994). Smith and Spiegelhalter (1980), Smith and Kohn (1996) and George and Foster (1997) adopted the prior

$$\boldsymbol{\beta}_{(m)}|\sigma^2, m \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, c^2\boldsymbol{V}_{(m)}\sigma^2\right). \tag{3.8}$$

with

$$\boldsymbol{\mu}_{\beta_{(m)}} = \mathbf{0}_{d(m)} \text{ and } \boldsymbol{V}_{(m)} = \left(\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)}\right)^{-1} \tag{3.9}$$

resulting to simple posterior odds with nice properties and interpretation. The prior parameter c fully controls the dimensionality penalty imposed to the log-maximum posterior distribution. For more details see Fernandez *et al.* (1998) and Chapter 6. For the specification of the prior parameter c^2, Smith and Kohn (1997) proposed values between 10 and 1000 while the values $c^2 = 100$ and $c^2 = n$ were highly recommended as good practical solutions. We argue that these choices work well in practice while the choice $c^2 = n - 1$ directly results in a Bayes factor which is equivalent to Bayes information criterion; see Chapter 6.

Ibrahim and Laud (1994) and Laud and Ibrahim (1995) used similar priors in their predictive model selection approach. They used covariance matrix (3.9) and mean $\boldsymbol{\mu}_{\beta_{(m)}}$ equal to the maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{(m)}$.

For the rest of generalised linear models we may use a more general form of prior distribution given by

$$\boldsymbol{\beta}_{(m)}|m \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, c^2\boldsymbol{V}_{(m)}\right). \tag{3.10}$$

As mentioned earlier, a usual choice, when no information is available, is $\boldsymbol{\mu}_{\beta_{(m)}} = \mathbf{0}_{d(m)}$. For the covariance matrix we may also consider the choice

$$\boldsymbol{V}_{(m)} = \left(\boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}\right)^{-1} \tag{3.11}$$

or alternatively use the inverse of the Fisher information matrix given by

$$\mathcal{I}_{\boldsymbol{\beta}_{(m)}} = -\left\{E\left[\frac{\partial^2 l(\boldsymbol{\beta}_{(m)})}{\partial\beta_{\nu,(m)}\partial\beta_{j,(m)}}\right]\right\}^{-1} = \left(\boldsymbol{X}_{(m)}^T \boldsymbol{H}_{(m)} \boldsymbol{X}_{(m)}\right)^{-1},$$

where

$$\boldsymbol{H}_{(m)} = Diag(h_i),\ h_i = \left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \frac{1}{a_i(\phi)b''(\vartheta)} = \{g'(E[Y_i])^2 a_i(\phi) v(E[Y_i])\}^{-1}, \tag{3.12}$$

$l\left(\boldsymbol{\beta}_{(m)}\right)$ is the likelihood function of model m. Alternatively, the above matrix can be substituted by the inverse of the observed information matrix given by

$$\boldsymbol{V}_{(m)} = \mathcal{I}_{\hat{\boldsymbol{\beta}}_{(m)}} = -\left[\frac{\partial^2 l(\boldsymbol{\beta}_{(m)})}{\partial\beta_{\nu,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\hat{\boldsymbol{\beta}}_{(m)}} = \left(\boldsymbol{X}_{(m)}^T \hat{\boldsymbol{H}}_{(m)} \boldsymbol{X}_{(m)}\right)^{-1} \tag{3.13}$$

where $\hat{\boldsymbol{\beta}}_{(m)}$ is the maximum likelihood estimate of $\boldsymbol{\beta}_{(m)}$. Although, prior distributions using a covariance resulting from (3.13) are data dependent, they can be thought as a non-informative since they do not influence the posterior $f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})$ for large values of c^2. Additionally, Smith and Kohn (1996) prior is a special case of the above generalization since $h_i = \sigma^{-2}$ for normal models.

Model	Link	GLM weights h_i
Normal	Identity	σ^{-2}
Poisson	Log	λ_i
Binomial	Logit	$N_i p_i(1-p_i)$
	Probit	$N_i/[p_i(1-p_i)\varphi^2(p_i)]$
	clog-log	$-N_i(1-p_i)log^2(1-p_i)/p_i$

Table 3.1: Generalised Linear Model Weights h_i.

Special case of prior (3.8) with $\boldsymbol{V}_{(m)}$ given by (3.13) is the *unit information prior* (for $c^2 = n$) . This prior has precision approximately equal to the precision provided by one data point. Fisher information matrix measures the amount of information provided by the data and therefore the precision of one data point is approximately given by $n^{-1}\mathcal{I}^{-1}_{\boldsymbol{\beta}_{(m)}}$. More detailed discussion of this prior is given by Spiegelhalter and Smith (1980), Kass and Wasserman (1995) and Pauler (1998). Pauler *et al.* (1998) use the unit root prior for model selection in variance component models.

3.2.1.3 Prior Distributions on the Coefficients Resulted from the Model with Standardised Variables

Raftery (1996a) developed a prior for generalised linear models with one dimensional terms by using standardised variables noted by Y^s and X^s_j. He initially considered the case where h_i is constant over all observations and identity link $g(\mu) = \mu$. If we consider the parameter vector of the full model $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ where β_0 is the coefficient for the constant term, then the new transformed model is given by

$$\mu^s_i = E(Y^s_i) = \beta^s_0 + \sum_{j \in \mathcal{V}(m)\setminus\{0\}} x^s_{ij}\beta^s_j.$$

Raftery (1996a) used independent normal priors on β^s_0 and β^s_j given by

$$\beta^s_0 \sim N(\mu_{\beta^s_0}, c_0^2), \quad \beta^s_j \sim N(0, c^2),$$

where c_0^2 and c^2 are prior parameters to be specified. From the above it is clear that

$$\beta_0 = \bar{y} + s_y\beta^s_0 - \sum_{j=1}^{p-1} \frac{s_y}{s_j}\bar{x}_j\beta^s_j, \quad \beta_j = \frac{s_y}{s_j}\beta^s_j,$$

where $\bar{y}$ and s_y^2 are the sample mean and variance of the response variable Y, $\bar{x}_j$ and s_j^2 are the sample mean and variance of the regressor X_j. This leads to a multivariate normal prior for the original parameters $\boldsymbol{\beta}$ given by (3.3) with prior mean

$$\boldsymbol{\mu}^T_{\beta_{(m)}} = (\mu_{\beta^s_0} + \bar{y}, 0, \ldots, 0)$$

and prior covariance matrix given by

$$\boldsymbol{\Sigma}_{(m)} = c^2 s_y^2 \begin{bmatrix} c^{-2}c_0^2 - \sum s_j^{-2}\bar{x}_j^2 & -s_2^{-2}\bar{x}_2 & -s_3^{-2}\bar{x}_3 & \ldots & -s^{-2}_{d(m)-1}\bar{x}_{d(m)-1} \\ -s_2^{-2}\bar{x}_2 & s_2^{-2} & 0 & \ldots & 0 \\ -s_3^{-2}\bar{x}_3 & 0 & s_3^{-2} & \ldots & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ -s^{-2}_{d(m)-1}\bar{x}_{d(m)-1} & 0 & \ldots & 0 & s^{-2}_{d(m)-1} \end{bmatrix}$$

where $d(m)$ is the dimension of the model (and, when we use one dimensional regressors, the number of covariates in model m).

For generalised linear models with other link functions he suggested similar procedure as above but the sample statistics should be weighted by h_i as defined in (3.12). Raftery (1996a) used two criteria (desiderata) to define plausible values for the parameter c^2. He reports the value of $c = 1.65$ as a trade-off between the two criteria and suggests various values from one to five ($1 \leq c \leq 5$).

A similar prior is defined by Raftery *et al.* (1997) for normal linear models. The prior for the constant term is given by

$$\beta_0 \sim N(\hat{\beta}_0, s_y^2 \sigma^2)$$

while for the other terms by

$$\beta_j \sim N(0, c^2 s_j^{-2} \sigma^2)$$

when they are continuous and by

$$\boldsymbol{\beta}_j \sim N(0, nc^2 \left(\boldsymbol{X}_j^T \boldsymbol{X}_j\right)^{-1} \sigma^2)$$

when they are categorical factors and $\boldsymbol{X}_j$ is the design matrix with each dummy variable centered on its sample mean. Raftery *et al.* (1997) following some specific criteria suggest to use prior values $c = 2.85$, $a_\tau = 1.27$ and $b_\tau = 0.3612$. They also argue that this prior corresponds to a Bayesian subjective prior and can be considered as an approximation of this true subjective prior.

3.2.1.4 Defining a Prior on the Full Model

Another strategy for the specification of prior distributions is to use a prior on the parameters $\boldsymbol{\beta}$ of the full model and the resulting marginal distributions for models of lower dimension. This approach was used by Kuo and Mallick (1998) but we argue that a prior on the full model may result in inappropriate or undesirable prior distributions for the models of lower dimension. In cases of block diagonal prior covariance matrix the prior distribution is decomposed to several independent prior distributions for the parameter vector $\boldsymbol{\beta}_j$ of each term j.

Following the logic of (3.8), we may use

$$\boldsymbol{\beta} \sim N\left(\mathbf{0}, c^2 \boldsymbol{V}\right).$$

Choices for the $\boldsymbol{V}$ are either $\boldsymbol{V} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ or $\boldsymbol{V} = (\boldsymbol{X}^T\hat{\boldsymbol{H}}\boldsymbol{X})^{-1}$ where $\boldsymbol{X}$ and $\hat{\boldsymbol{H}}$ are the design (or data) and weight matrices of the full model respectively. Note that such prior on Poisson log-linear and logistic regression models with categorical regressors and sum-to-zero constraints leads to the distribution of Dellaportas and Forster (1999) given by (3.5). In normal models the above prior can also be defined conditionally on σ^2 similar to Smith and Kohn (1996) prior.

3.2.1.5 Intrinsic, Conjugate and Imaginary Samples Prior Distributions

Another more complicated approach is to consider training samples. Training samples can be either a subset of our dataset or an imaginary sample used to express our prior beliefs. Spiegelhalter and Smith (1982) favour the idea of imaginary training samples. Elicitation of prior beliefs for variable selection in normal models are given by Garthwaite and Dickey (1992) and Ibrahim and Chen (1999). Berger and Pericchi (1996a) introduced the intrinsic Bayes factor (see also Section 2.2.5) based on the idea of using a minimal part of the data for estimation while the rest for model selection. They argue that intrinsic Bayes factors correspond to actual Bayes factor for a sensible prior. These priors are called intrinsic priors.

Alternative approaches include the use of conjugate priors on canonical parameters ϑ_i proposed in minimal Kullback-Leibler approach by Goutis and Robert (1998).

Finally, according to a different perspective, we may utilize the idea of Bedrick *et al.* (1996) for any model selection problem. For example in the binomial models with different link functions a beta prior on binomial probabilities may be used and then calculate the prior distribution of the model coefficients. Using a beta prior on each p_i with parameters a_{p_i} and b_{p_i}, assuming that the inverse of the design matrix of the full (saturated) model $\boldsymbol{X}$ exists, results in

$$f(\boldsymbol{\beta}) = \frac{\Gamma(a_{p_i} + b_{p_i})}{\Gamma(a_{p_i})\Gamma(b_{p_i})} \prod_{i=1}^{p} \frac{[exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{a_{p_i}}}{[1 + exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{a_{p_i}+b_{p_i}}} |\mathbf{X}| \tag{3.14}$$

for the logit link. The corresponding prior for the probit link is given by

$$f(\boldsymbol{\beta}) = \frac{\Gamma(a_{p_i} + b_{p_i})}{\Gamma(a_{p_i})\Gamma(b_{p_i})} \prod_{i=1}^{p} \frac{[exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{a_{p_i}-1}}{[1 + exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{a_{p_i}+b_{p_i}-2}} \varphi(\boldsymbol{x}_i^T\boldsymbol{\beta})|\mathbf{X}| \tag{3.15}$$

where $\varphi(.)$ is the density function of the standard normal distribution. Finally, the density

$$f(\boldsymbol{\beta}) = \frac{\Gamma(a_{p_i}+b_{p_i})}{\Gamma(a_{p_i})\Gamma(b_{p_i})} \prod_{i=1}^{p} \frac{[exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{a_{p_i}}}{[1+exp(\boldsymbol{x}_i^T\boldsymbol{\beta})]^{a_{p_i}+b_{p_i}-2}} exp\left(-e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)|\mathbf{X}| \tag{3.16}$$

defines the prior distribution for the complementary log-log link. The above priors are defined for the full (saturated) model while all other priors are given by the corresponding marginal distributions. Although these priors seem plausible, they do not solve the problem of prior specification since a uniform prior on binomial probabilities strongly supports the constant model and it unnecessarily complicates the conditional posterior distributions.

3.2.2 Prior Distribution on Model Space

The problem is not only to define the prior on model space $\mathcal{M}$ but also to decide how large $\mathcal{M}$ should be. Given that we restrict attention to a limited set of model $\mathcal{M}$, the uniform distribution on this model space is frequently used as 'non-informative' prior because it gives the same weight in all models included in model M. Therefore, we have

$$f(m) = \frac{1}{|\mathcal{M}|},\ \forall\ m \in \mathcal{M}. \tag{3.17}$$

Suppose that the set of all possible models is $\mathcal{M}^*$. The selection of $\mathcal{M} \subset \mathcal{M}^*$ and assignment of uniform distribution on this subset is totally misleading and is equivalent to consider $\mathcal{M}^*$ and set the prior probability of all models $m \in \mathcal{M}^* \setminus \mathcal{M}$ equal to zero. Under this condition there is not any prior distribution that can be assumed as non-informative unless the set of all possible models can be exactly determined; for further discussion on this very important issue see Draper (1995).

In variable selection, when no restrictions on variable combinations are imposed, each model prior can be easily decomposed to independent Bernoulli distributions for each term indicator γ_j

$$\gamma_j \sim Bernoulli(\pi_j),\ j \in \mathcal{V}, \tag{3.18}$$

where π_j is the prior probability to include j term in the model. In the non-informative case, the uniform prior on $\mathcal{M}$ corresponds to $\pi_j = 0.5$ for all $j \in \mathcal{V}$. The above prior can also be written in the form

$$f(\boldsymbol{\gamma}) = \prod_{j\in\mathcal{V}} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j}.$$

Usually we consider the same prior probability for all terms under consideration, that is $\pi_j = \pi$, for all $j \in \mathcal{V}$ resulting to

$$f(\boldsymbol{\gamma}) = \pi^{d(\boldsymbol{\gamma})}(1-\pi)^{p-d(\boldsymbol{\gamma})}$$

where $d(\boldsymbol{\gamma})$ is the dimension of $\boldsymbol{\gamma}$ model, hence $d(\boldsymbol{\gamma}) = \sum_{j \in \mathcal{V}} \gamma_j$. It is straightforward that

$$f(\boldsymbol{\gamma}) \propto \left(\frac{\pi}{1-\pi}\right)^{d(\boldsymbol{\gamma})} = [PrO]^{d(\boldsymbol{\gamma})}$$

which denotes that the prior probability of a model depends on its dimension and parameter PrO which measures the prior odds of including any term in the model equation.

When restrictions on model space are imposed (for example selection of hierarchical models only in contingency tables) the prior on $\boldsymbol{\gamma}$ needs to be specified hierarchically. Chipman (1996) demonstrates how we can define conditional probabilities in order to achieve the prior (3.17). He also allowed to visit non-hierarchical models with low probability.

Using similar ideas we argue that the probability of a model should be written as a product of conditional distributions of term indicators and therefore it can be generally expressed either as

$$f(\boldsymbol{\gamma}) = \prod_{j \in \mathcal{V}} f(\gamma_j | \{\gamma_k : j \in \mathcal{V}(k)\})$$

or

$$f(\boldsymbol{\gamma}) = \prod_{j \in \mathcal{V}} f(\gamma_j | \{\gamma_k : k \in \mathcal{V}(j)\}).$$

For example, the probability of a three way ANOVA model with $\mathcal{V} = \{A, B, C, AB, AC, BC, ABC\}$ can be written as

$$f(\boldsymbol{\gamma}) = f(\gamma_{ABC})f(\gamma_{AB}, \gamma_{AC}, \gamma_{BC}|\gamma_{ABC})f(\gamma_A|\gamma_{AB}, \gamma_{AC})f(\gamma_B|\gamma_{AB}, \gamma_{BC})f(\gamma_C|\gamma_{AC}, \gamma_{BC})$$

or

$$f(\boldsymbol{\gamma}) = f(\gamma_A)f(\gamma_B)f(\gamma_C)f(\gamma_{AB}|\gamma_A, \gamma_B)f(\gamma_{AC}|\gamma_A, \gamma_C)f(\gamma_{BC}|\gamma_B, \gamma_C)f(\gamma_{ABC}|\gamma_{AB}, \gamma_{AC}, \gamma_{BC}).$$

Let us consider the simple example of a two way contingency table with five hierarchical models ($|\mathcal{M}| = 5$), $\mathcal{M} = \{\emptyset, [A], [B], [A][B], [AB]\}$ and $\mathcal{V} = \{A, B, AB\}$; where $\emptyset$ is the constant model. initially, we set $f(\gamma_{AB}) = 1/5$ and then $f(\gamma_A|\gamma_{AB}) = f(\gamma_B|\gamma_{AB}) = (1/2)^{1-\gamma_{AB}}$ resulting in $f(\boldsymbol{\gamma}) = 1/5$ for all $\boldsymbol{\gamma}$ under consideration.

The prior distributions (3.17) and (3.18) with $\pi_j = 0.5$ (or $PrO = 1.0$) are widely used as non-informative priors since they give the same prior weight to all models. George and Foster (1997) argue that such prior distributions give more weight to models with dimension close to $p/2$ and hence are informative in terms of dimensionality. They alternatively propose the specification of variable probabilities using empirical Bayes procedures. On the other hand, Laud and Ibrahim (1996) propose predictive methods for the specification of prior model probabilities.

An alternative approach includes specification of prior model probabilities depending on the prior precision of model parameters $\boldsymbol{\beta}$. For this reason, the prior parameter PrO, included in $f(m)$, should be defined as a function of and prior covariance $\boldsymbol{\Sigma}_{(m)}$ or the parameter controlling the flatness of the prior distribution, c^2. For further details and new developments see Chapter 6.

3.2.3 An Alternative Prior Specification

An alternative approach can be used for specifying the prior distribution $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma})$. Instead of specifying $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma})$ and $f(\boldsymbol{\gamma})$ we may consider the possibility, at least in variable selection, to specify $f(\boldsymbol{\gamma}|\boldsymbol{\beta})$ and $f(\boldsymbol{\beta})$. Consider the simple case of $y_i \sim N(x_i\gamma\beta, \sigma^2)$. We want to test whether β is significant or not, that is to estimate $f(\gamma)$. The prior $f(\gamma = 1|\beta)$ will the probability of inclusion of X under specific values of β. For example, the value β^s such that $f(\gamma = 1|\beta^s) = 0.5$ will be considered as the value that we are a-priori totally uncertain whether we should include or not X in our model. Although this prior has nice interpretation in variable selection problems, it has major drawbacks. The main drawback is that such prior distributions will complicate the posterior distributions and the sampling procedures. A further drawback is that we do not know what choices of $f(\gamma|\beta)$ are plausible and what effect will have on the MCMC methods. Moreover, it is quite unclear what distribution will be appropriate for $f(\beta)$. A suggested prior, that needs further exploration, is given by

$$f(\gamma_j = 1|\beta_j) = \frac{\xi_1 e^{\beta_j}}{\xi_0 + \xi_1 e^{\beta_j}}$$

where ξ_0/ξ_1 is the prior odds to exclude j term from the model when all parameters are zero. This interpretation closely related with the interpretation of parameter k_j used in SSVS; see

Section 3.4.1. Another possible choice is

$$f(\gamma_j = 1|\beta_j) = \frac{\xi_1 \beta_j^2}{\xi_0 + \xi_1 \beta_j^2}$$

resulting $f(\gamma_j = 1|\beta_j = 0) = 0$ so $f(\gamma_j = 1|\beta_j = \pm\sqrt{\xi_0/\xi_1}) = 0.5$. A logical choice for the latter is $\sqrt{\xi_0/\xi_1} = 2\sqrt{Var(\hat{\beta}_j)}$; however we have not pursued this issue further.

3.3 MCMC Model Selection Methods

3.3.1 Reversible Jump

Reversible jump (Green, 1995) is a flexible MCMC sampling strategy for generating observations from the joint posterior distribution $f(m, \boldsymbol{\beta}_{(m)}|\boldsymbol{y})$. The method is based on creating a Markov chain which can 'jump' between models with parameter spaces of different dimension, while retaining detailed balance which ensures the correct limiting distribution, provided the chain is irreducible and aperiodic.

Suppose that the current state of the Markov chain is $\left(m, \boldsymbol{\beta}_{(m)}\right)$, where $\boldsymbol{\beta}_{(m)}$ has dimension $d(m)$, then one version of the procedure is as follows

- Propose a new model m' with probability $j(m, m')$.
- Generate $\boldsymbol{u}$ from a specified proposal density $q(\boldsymbol{u}|\boldsymbol{\beta}_{(m)}, m, m')$.
- Set $(\boldsymbol{\beta}'_{(m')}, \boldsymbol{u}') = h_{m,m'}(\boldsymbol{\beta}_{(m)}, \boldsymbol{u})$ where $h_{m,m'}$ is a specified invertible function. Hence $d(m) + d(\boldsymbol{u}) = d(m') + d(\boldsymbol{u}')$. Note that $h_{m',m} = h^{-1}_{m,m'}$.
- Accept the proposed move to model m' with probability

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{(m')}, m') f(\boldsymbol{\beta}'_{(m')}|m') f(m') j(m', m) q(\boldsymbol{u}'|\boldsymbol{\beta}'_{(m')}, m', m)}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|m) f(m) j(m, m') q(\boldsymbol{u}|\boldsymbol{\beta}_{(m)}, m, m')} \left|\frac{\partial h(\boldsymbol{\beta}_{(m)}, \boldsymbol{u})}{\partial(\boldsymbol{\beta}_{(m)}, \boldsymbol{u})}\right|\right). \tag{3.19}$$

There are many variations or simpler versions of reversible jump that can be applied in specific model selection problems. In particular, if all parameters of the proposed model are generated from a proposal distribution, then $(\boldsymbol{\beta}'_{(m')}, \boldsymbol{u}') = (\boldsymbol{u}, \boldsymbol{\beta}_{(m)})$ with $d(m) = d(\boldsymbol{u}')$ and $d(m') = d(\boldsymbol{u})$, and the Jacobian term in (3.19) is one. This version of reversible jump could be used for jumping between models for which no appropriate parameter transformation exists.

Where models m and m' may be described as nested, then there may be an extremely natural proposal distribution and transformation function $h_{m,m'}$ (may be the identity function) such that $d(\boldsymbol{u}') = 0$ and $\boldsymbol{\beta}_{m'} = h_{m,m'}(\boldsymbol{\beta}_{(m)}, \boldsymbol{u})$. See, for example, Dellaportas and Forster (1999). Finally, if $m' = m$, then the move is a standard Metropolis-Hastings step.

3.3.2 Carlin and Chib's Method

Carlin and Chib (1995) proposed using a Gibbs sampler to generate from the posterior distribution $f(m, \boldsymbol{\beta}_{(m)}|\boldsymbol{y})$. In order to do this, it is required to consider a Markov chain of realisations of $\{m, \boldsymbol{\beta}_{(m_k)} : m_k \in \mathcal{M}\}$. Therefore, a prior distribution for $\{m, \boldsymbol{\beta}_{(m_k)} : m_k \in \mathcal{M}\}$ is no longer completely specified by $f(m)$ and $f(\boldsymbol{\beta}_{(m)}|m)$, so Carlin and Chib proposed the use of pseudopriors or linking densities $f(\boldsymbol{\beta}_{(m_k)}|m \neq m_k)$, $m_k \in \mathcal{M}$.

The full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_{(m_k)}|\boldsymbol{y},\ \{\boldsymbol{\beta}_{(m_l)} : m_l \neq m_k\}, m) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m) f(\boldsymbol{\beta}_{(m)}|m) & m_k = m \\ f(\boldsymbol{\beta}_{(m_k)}|m_k \neq m) & m_k \neq m \end{cases} \tag{3.20}$$

where $\{\boldsymbol{\beta}_{(m_l)} : m_l \neq m_k\}$ are the parameter vectors $\boldsymbol{\beta}_{(m_l)}$ for all $m_l \in \mathcal{M} \setminus \{m_k\}$. The full conditional posterior distribution of the model indicator is given by

$$f(m|\{\boldsymbol{\beta}_{(m_k)} : m_k \in \mathcal{M}\}, \boldsymbol{y}) = \frac{A_m}{\sum\limits_{m_k \in \mathcal{M}} A_{m_k}} \tag{3.21}$$

where $\{\boldsymbol{\beta}_{(m_k)} : m_k \in \mathcal{M}\}$ are the parameter vectors $\boldsymbol{\beta}_{(m_k)}$ for all $m_k \in \mathcal{M}$ and

$$A_m = f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m) \prod_{m_l \in \mathcal{M}} \left\{ f(\boldsymbol{\beta}_{(m_l)}|m) \right\} f(m).$$

When $m_k = m$, $\boldsymbol{\beta}_{(m_k)}$ are generated from the conditional posterior distribution $f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})$, and when $m_k \neq m$ from the corresponding pseudoprior, $f(\boldsymbol{\beta}_{(m_k)}|m)$. Since for $m_k \neq m$ we generate directly from the pseudopriors $f(\boldsymbol{\beta}_{(m_k)}|m)$, Carlin and Chib sampler will be optimal when these densities are good approximations of the conditional posterior distributions $f(\boldsymbol{\beta}_{(m_k)}|m_k, \boldsymbol{y})$ and therefore we only need one density $f(\boldsymbol{\beta}_{(m_k)}|m)$ for all $m \in \mathcal{M} \setminus \{m_k\}$. In the following we denote this common pseudoprior as $f(\boldsymbol{\beta}_{(m_k)}|m_k \neq m)$ for all $m \in \mathcal{M} \setminus \{m_k\}$. The model indicator m is generated as a discrete random variable.

The main drawback of this method is the unavoidable specification of, and generation from, many pseudoprior distributions. Carlin and Chib (1995) point out that, pseudopriors should be chosen to make the method efficient, since they do not enter the marginal posterior distributions $f\left(m, \boldsymbol{\beta}_{(m)}|\boldsymbol{y}\right)$ of interest. However, generation from $|\mathcal{M}|-1$ pseudopriors at every cycle of the Gibbs sampler is still required, and this is computationally demanding.

3.3.3 Markov Chain Monte Carlo Model Composition (MC^3)

Markov chain Monte Carlo model composition (MC^3) was introduced by Madigan and York (1995) in graphical model selection. Variants of MC^3 were used in normal linear models by Hoeting *et al.* (1995, 1996), Raftery *et al.* (1997). MC^3 is a simple Metropolis algorithm which helps to explore the model space when the number of candidate models is large. We define as neighbourhood of model m the set $nb(m)$ which includes all models that differ from m by one term or variable. We also select a transition function $j(m, m')$ for all $m, m' \in \mathcal{M}$ which indicates the probability of proposing model m' when we are currently in model m. Note that $j(m, m') = |nb(m)|^{-1}$, for all $m' \in nb(m)$ and $j(m, m') = 0$, for all $m' \notin nb(m)$; where $|nb(m)|$ are the number of models in $nb(m)$. If the chain is currently in state m then we propose model m' with probability $j(m, m')$ and accept this proposed model with probability

$$\alpha = min\left(1, \frac{f(m'|\boldsymbol{y})|nb(m)|}{f(m|\boldsymbol{y})|nb(m')|}\right).$$

The above procedure composes the definition of MC^3 for graphical model selection by Madigan and Raftery (1994). In the case where $|nb(m)| = |nb(m')|$ the above acceptance probability is simplified to $\alpha = min(1, PO_{m',m})$, as given by Kass and Raftery (1995), Madigan *et al.* (1995) and Raftery *et al.* (1997). We can easily generalise MC^3 by using any $j(m, m')$ and therefore accept the proposed model m'

$$\alpha = min\left(1, \frac{f(m'|\boldsymbol{y})j(m', m)}{f(m|\boldsymbol{y})j(m', m)}\right). \tag{3.22}$$

In cases at which the posterior odds (or Bayes factor) cannot be calculated analytically, BIC or Laplace approximations may be used instead. The Metropolis step, when we are in model m and propose to switch to model m', is given by

$$\alpha = min\left(1, \frac{|\boldsymbol{X}^T_{(m')}\hat{\boldsymbol{H}}_{(m')}\boldsymbol{X}_{(m')}|^{\frac{1}{2}} f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m')}, m') f(\hat{\boldsymbol{\beta}}_{(m')}|m') j(m', m)}{|\boldsymbol{X}^T_{(m)}\hat{\boldsymbol{H}}_{(m)}\boldsymbol{X}_{(m)}|^{\frac{1}{2}} f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m)}, m) f(\hat{\boldsymbol{\beta}}_{(m)}|m) j(m, m')} (2\pi)^{[d(m')-d(m)]/2}\right)$$

when Laplace approximation is adopted or

$$\alpha = min\left(1, \frac{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m')}, m')j(m', m)}{f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m)}, m)j(m, m')} n^{-[d(m')-d(m)]/2}\right)$$

is based on BIC approximation.

The results from approximate MC^3 samplers can be used as a yardstick for further analysis or as proposal distributions in more advanced MCMC model selection algorithms such as reversible jump and Carlin and Chib sampler. These approximation should be handled with care since regularity conditions must hold; for details see Pauler (1998). Laplace approximations should be preferred since it provides more accurate results and allows for prior adjustment.

3.4 Variable Selection

As we have already mentioned, in variable selection problems statistical models may be represented naturally as $(s, \boldsymbol{\gamma}) \in \mathcal{S} \times \{0,1\}^p$, where the indicator vector $\boldsymbol{\gamma}$ represents which of the p possible sets of covariates are present in the model and s represents other structural properties of the model. For example, in generalised linear models, s may describe the distribution, link function and variance function, and the linear predictor is given by (3.2). In the following, we restrict consideration to variable selection aspects assuming that s is known, or dealt with in another way and therefore we substitute $\boldsymbol{\gamma}$ for model indicator m. For example, we can apply reversible jump to variable selection by substituting $\boldsymbol{\gamma}$ for m in (3.19).

3.4.1 Stochastic Search Variable Selection

Stochastic Search Variable Selection (SSVS) was introduced by George and McCulloch (1993) for linear regression models and has been adopted for more complicated cases in pharmacokinetics, finance, generalised linear models, log-linear models and multivariate regression models.

3.4.1.1 The Method

The difference between SSVS and other variable selection approaches is that the parameter vector $\boldsymbol{\beta}$ is specified to be of full dimension p under all models, so the linear predictor is $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ instead of (3.2) for all models, where $\boldsymbol{X}$ contains all the potential explanatory variables. The indicator variables γ_j are involved in the modelling process through the prior

$$\boldsymbol{\beta}_j|\gamma_j \sim \gamma_j N(0, \boldsymbol{\Sigma}_j) + (1-\gamma_j)N(0, k_j^{-2}\boldsymbol{\Sigma}_j) \tag{3.23}$$

for specified k_j and $\boldsymbol{\Sigma}_j$. The prior parameters k_j and $\boldsymbol{\Sigma}_j$ in (3.23) are chosen so that when $\gamma_j = 0$ (covariate is 'absent' from the linear predictor) the prior distribution for $\boldsymbol{\beta}_j$ ensures that $\boldsymbol{\beta}_j$ is constrained to be 'close to $\mathbf{0}$'. When $\gamma_j = 1$ the prior is diffuse, assuming that little prior information is available about $\boldsymbol{\beta}_j$.

The full conditional posterior distributions of $\boldsymbol{\beta}_j$ and γ_j are given by

$$f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_j|\gamma_j) \tag{3.24}$$

and

$$\frac{f(\gamma_j = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = \frac{f(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}|\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \tag{3.25}$$

where $\boldsymbol{\gamma}_{\setminus j}$ denotes all terms of $\boldsymbol{\gamma}$ except γ_j.

If we use the prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ defined by (3.23) and assume that $f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j}) = f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j}) = |\mathcal{M}|^{-1}$ for all $j \in \mathcal{V}$, then

$$\frac{f(\gamma_j = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = k_j^{-d_j} exp\left(\frac{k_j^2 - 1}{2} \boldsymbol{\beta}_j^T \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\beta}_j\right). \tag{3.26}$$

3.4.1.2 Priors for Stochastic Search Variable Selection

The posterior model probabilities are heavily dependent on the choice of the prior parameters k_j^2 and $\boldsymbol{\Sigma}_j$. One way of specifying these parameters is by setting $\boldsymbol{\Sigma}_j$ as a diffuse prior (for $\gamma_j = 1$) and then choosing k_j^2 by considering the value of $|\boldsymbol{\beta}_j|$ at which the densities of the two components of the prior distribution are equal. This can be considered to be the smallest value of $|\boldsymbol{\beta}_j|$ at which the term is considered of practical significance.

When β_j is one-dimensional then specification of the prior may be completed by using the methodology of George and McCulloch (1993), using the value of $|\beta_j|$ at which the densities of the two components of the prior distribution are equal. In this case $f(\gamma_j = 0|\beta_j) = f(\gamma_j =$

$1|\beta_j)$, and therefore it may be considered as the smallest value for which the term is thought to be of practical significance. Now, suppose that δ_j is the smallest value of β_j of practical significance. Then

$$\delta_j \;=\; \sqrt{\Sigma_j \frac{2\log k_j}{(k_j^2-1)}} \;\approx\; \sqrt{\Sigma_j \frac{2\log k_j}{k_j^2}}, \tag{3.27}$$

where Σ_j is the prior variance when j is included in the model. From the above we may specify δ_j and then try to identify optimal selections of Σ_j and k_j. In some cases, the user defines the 'large' variance (Σ_j) and may use the above equation to define the 'small' variance. Moreover, $\delta_j k_j \Sigma_j^{-1/2}$ changes slowly with variations of k_j since for k_j =10, 100, 1,000, 10,000, 100,000 then $\delta_j k_j \Sigma_j^{-1/2} = 2.1,\ 3.1,\ 3.7,\ 4.3,\ 4.8$. The semiautomatic approach considers the two marginal distributions $\hat{\beta}_j|\sigma^2_{\hat{\beta}_j}, \gamma_j = 0 \sim N\left(0, \sigma^2_{\hat{\beta}_j} + k_j^{-2}\Sigma_j\right)$ and $\hat{\beta}_j|\sigma^2_{\hat{\beta}_j}, \gamma_j = 1 \sim N\left(0, \sigma^2_{\hat{\beta}_j} + \Sigma_j\right)$, where $\sigma^2_{\hat{\beta}_j}$ is the variance of the least square estimates $\hat{\beta}_j$ and their intersection point; for more details see George and McCulloch (1993).

According to George and McCulloch (1997), SSVS gives results close to the actual posterior probabilities for large values of k_j. On the other hand, the largest the prior parameter k_j the slowest the convergence of the chain. Hence, we propose to specify Σ_j as described in usual Bayesian model selection methods and k_j in such way that gives results close to the actual posterior probabilities and also does not prevent the chain to converge in reasonable time. We report that $k_j = 1000$ is a sensible choice. Moreover,

$$\frac{f(\gamma_j = 0|\beta_j = 0)}{f(\gamma_j = 1|\beta_j = 0)} = k_j \frac{f(\gamma_j = 0)}{f(\gamma_j = 1)}.$$

and therefore k_j can be interpreted as the prior odds that j term should be excluded from the model if β_j is zero and $f(\gamma_j = 1) = 1/2$. Under this interpretation the values $100 < k_j < 1000$ seem plausible choices. We can still use (3.27) to monitor the area of non-significant values for different choices of priors. Similar methods are proposed in Chapter 4 for the specification of prior distributions in log-linear interaction models where $\boldsymbol{\beta}_j$ terms are multidimensional.

George and McCulloch (1993) also proposed an alternative prior set-up based on a multivariate normal distribution. This prior is given by

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N(\mathbf{0}, \boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{R}_{\boldsymbol{\gamma}}\boldsymbol{D}_{\boldsymbol{\gamma}}), \quad \boldsymbol{D}_{\boldsymbol{\gamma}} = diag[k_j^{\gamma_j - 1}\Sigma_j^{1/2}] \tag{3.28}$$

where $\boldsymbol{R}_{\boldsymbol{\gamma}}$ is the prior correlation matrix and Σ_j is the prior variance when the j term is in the model. George and McCulloch (1993) propose $\boldsymbol{R}_{\boldsymbol{\gamma}} = \mathbf{I}$ and $\boldsymbol{R}_{\boldsymbol{\gamma}} \propto (\boldsymbol{X}^T\boldsymbol{X})^{-1}$. We may generalise the latter proposed prior correlation for all generalised linear models by using $\boldsymbol{R}_{\boldsymbol{\gamma}} \propto \boldsymbol{\mathcal{I}}_{\hat{\boldsymbol{\beta}}}$; for further details see George and McCulloch (1997).

3.4.2 Kuo and Mallick Variable Selection

Kuo and Mallick (1998) advocated the use of the linear predictor (3.2) for variable selection. They considered a prior distribution $f(\boldsymbol{\beta})$ which is independent of $\boldsymbol{\gamma}$ (and therefore m) so that $f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}) = f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j})$

Therefore, the full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}) & \gamma_j = 1 \\ f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}) & \gamma_j = 0 \end{cases} \tag{3.29}$$

and

$$\frac{f(\gamma_j = 1|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}. \tag{3.30}$$

The above approach is extremely straightforward. It only requires to specify the usual prior on $\boldsymbol{\beta}$ (for the full model) and the conditional prior distributions $f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j})$ replace the pseudopriors required by Carlin and Chib's method. However, this simplicity may also be a drawback, as there is no flexibility here to alter the method to improve efficiency. In practice, if, for any $\boldsymbol{\beta}_j$, the prior is diffuse compared with the posterior, the method may be inefficient.

3.5 Model Selection Methods for Linear Normal Models Using Marginal Posterior Distributions

This section critically reviews and generalises MCMC methodologies used to explore model space in normal models. In normal models we can exactly evaluate the posterior odds if we use the prior distributions (3.6) and (3.7) which result to the conjugate joint prior called normal inverse gamma. The only problem is such cases is the calculation of all posterior probabilities when the set of models is large. Therefore these methods sample directly from the target distribution $f(\boldsymbol{\gamma}|\boldsymbol{y})$, or $f(m|\boldsymbol{y})$, without the need to generate samples also from model parameters $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ and σ^2 or any other pseudo-parameters. For this reason we call these methods 'fast' model selection algorithms. Similar terminology was also used by

Chipman (1997). We divide this subsection into three parts. The first describes all the variable selection methods and their associations, the second describes the transformations advocated and the last is concerned with the outlier identification.

3.5.1 Fast Variable Selection Algorithms

In this section we consider the conjugate prior (3.6) and (3.7) resulting to a normal inverse gamma prior distribution for $\boldsymbol{\beta}_{(m)}$ and σ^2.

The fast variable selection methods involve generation of the model indicator m or $\boldsymbol{\gamma}$ directly from the marginal posterior distribution $f(m|\boldsymbol{y})$ or $f(\boldsymbol{\gamma}|\boldsymbol{y})$. The approach varies according to the sampler (Metropolis or Gibbs approach), the prior distributions used and the model indicator approach adopted (m or $\boldsymbol{\gamma}$). We will try to be as general as possible adopting the general normal inverse gamma prior setup and restrict attention to special cases presented by other authors.

Smith and Kohn (1996) developed a Gibbs sampler for variable selection for nonparametric regression in normal models. The resulted Gibbs sampler involves sequential generation of each γ_j from a Bernoulli distribution with success probability $O_j/(1+O_j)$; where O_j is given by

$$O_j = \frac{f(\gamma_j=1|\boldsymbol{\gamma}_{\backslash j},\boldsymbol{y})}{f(\gamma_j=0|\boldsymbol{\gamma}_{\backslash j},\boldsymbol{y})} =$$

$$= \left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})}||\boldsymbol{\Sigma}_{(\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})}|}{|\boldsymbol{\Sigma}_{(\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})}||\tilde{\boldsymbol{\Sigma}}_{(\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})}|}\right)^{1/2} \left(\frac{SS_{\gamma_j=1,\boldsymbol{\gamma}_{\backslash j}}+2b_\tau}{SS_{\gamma_j=0,\boldsymbol{\gamma}_{\backslash j}}+2b_\tau}\right)^{-n/2-a_\tau} \frac{f(\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})}{f(\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})}, \quad (3.31)$$

where a_τ and b_τ are prior parameters of the precision with usual choices of $a_\tau=b_\tau=0$, $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}$ is the posterior covariance matrix given by

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}} = \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{X}_{(\boldsymbol{\gamma})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\right)^{-1}$$

and $SS_{\boldsymbol{\gamma}}$ are the posterior residual sum of squares given by

$$SS_{\boldsymbol{\gamma}} = \boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\mu}^T_{\beta_{(\gamma)}}\boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}} - \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_\gamma}\right)^T \tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})} \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_\gamma}\right) \quad (3.32)$$

when the general normal inverse gamma prior setup is adopted. Smith and Kohn (1996) used the more restrictive prior (3.8) with prior parameters given by (3.9) which is related to

Zellner's g-priors; see Zellner (1986). The posterior distributions are now simplified to

$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = (c^2+1)^{-d_j/2} \left(\frac{SS_{\gamma_j=1,\boldsymbol{\gamma}_{\setminus j}} + 2b_\tau}{SS_{\gamma_j=0,\boldsymbol{\gamma}_{\setminus j}} + 2b_\tau} \right)^{-n/2-a_\tau} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}, \quad (3.33)$$

where where $SS_{\boldsymbol{\gamma}}$ are given by

$$SS_{\boldsymbol{\gamma}} = \boldsymbol{y}^T\boldsymbol{y} - \frac{c^2}{c^2+1}\boldsymbol{y}^T\boldsymbol{X}_{(\boldsymbol{\gamma})}\left(\boldsymbol{X}_{(\boldsymbol{\gamma})}^T\boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1}\boldsymbol{X}_{(\boldsymbol{\gamma})}^T\boldsymbol{y}. \quad (3.34)$$

Under the same prior setup Smith and Kohn (1996) type of Gibbs samplers are closely related to MC^3 for normal models provided that we substitute the model indicator m by $\boldsymbol{\gamma}$ and the Metropolis step by sequential Gibbs steps. One sampling step in MC^3 is equivalent to updating one γ_j using a Metropolised version of Smith and Kohn sampler. Therefore, the MC^3 step for accepting a proposed move from model m to model $m' \in nb(m')$ (model m' differs from model m only in j term) is given by

$$\alpha \;=\; \begin{cases} min\left(1, O_j \, \frac{j(m',m)}{j(m,m')}\right) & \text{when} \quad \gamma_j = 1 \\ min\left(1, O_j^{-1} \frac{j(m',m)}{j(m,m')}\right) & \text{when} \quad \gamma_j = 0 \end{cases}.$$

The ratio O_j will be given by equations (3.31) and (3.32) if the general normal inverse gamma setup is adopted or by equations (3.33) and (3.34) in the case of the Smith and Kohn prior setup.

Consider now two models $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}''$ differing only in the jth term, that is $\gamma_j' = 1$ and $\gamma_j'' = 0$ while $\gamma_\nu' = \gamma_\nu''$ for all $\nu \in \mathcal{V} \setminus \{j\}$. Then without loss of generality we set $\boldsymbol{X}_{(\boldsymbol{\gamma}')} = [\boldsymbol{X}_{(\boldsymbol{\gamma}'')}, \boldsymbol{X}_j]$. Under the above assumption, the posterior residual sum of squares with Smith and Kohn (1996) prior is given by

$$SS_{\boldsymbol{\gamma}'} = SS_{\boldsymbol{\gamma}''} + \frac{c^2}{c^2+1}\boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\gamma}''}\right)\left(\boldsymbol{X}_j\left(\boldsymbol{X}_j^T\boldsymbol{X}_j - \boldsymbol{X}_j^T\boldsymbol{P}_{\boldsymbol{\gamma}''}\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j^T\right)\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{\gamma}''}\right)\boldsymbol{y}$$

where

$$\boldsymbol{P}_{\boldsymbol{\gamma}} = \boldsymbol{X}_{(\boldsymbol{\gamma})}\left(\boldsymbol{X}_{(\boldsymbol{\gamma})}^T\boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1}\boldsymbol{X}_{(\boldsymbol{\gamma})}^T.$$

In orthogonal cases the posterior residual sum of squares (3.34) simplifies to

$$SS_{\boldsymbol{\gamma}} = \boldsymbol{y}^T\boldsymbol{y} - \frac{c^2}{c^2+1}\sum_{j\in\mathcal{V}}\gamma_j F_j$$

$$F_j = \boldsymbol{y}^T\boldsymbol{X}_j\left(\boldsymbol{X}_j^T\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j^T\boldsymbol{y} = \hat{\boldsymbol{\beta}}_j^T\boldsymbol{X}_j^T\boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j. \quad (3.35)$$

Substituting the above formula in the posterior distribution of variable indicators results in Clyde *et al.* (1996) Gibbs sampler for orthogonal cases with prior (3.9) and $a_\tau = b_\tau = 0$. The resulting conditional posterior term probabilities are given by

$$\frac{f(\gamma_j = 1|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = (c^2+1)^{-d_j/2}\left(1 - \frac{c^2}{c^2+1}F_j/SS_{\gamma_j=0,\boldsymbol{\gamma}_{\setminus j}}\right)^{-n/2}. \tag{3.36}$$

Clyde (1999) introduced a straightforward sampler for linear regression models with known variance. A more general version of Clyde (1999), ideal for ANOVA model selection with sum-to-zero constraints, is developed in this section. Clyde (1999) utilizes clever ideas similar to Foster and George (1994) where they use information criteria in orthogonal data to select variables rather than models.

Assuming known σ^2 and the prior distribution (3.4) the posterior conditional for γ_j will be (as usually) Bernoulli with success probability $O_j/(1+O_j)$ and O_j is given by

$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{y}, \sigma^2, \gamma_{\setminus j})}{f(\gamma_j = 0|\boldsymbol{y}, \sigma^2, \gamma_{\setminus j})} = \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}O'_j$$

$$O'_j = \left(\frac{|\boldsymbol{X}_j^T\boldsymbol{X}_j/\sigma^2 + \boldsymbol{\Sigma}_j^{-1}|}{|\boldsymbol{\Sigma}_j^{-1}|}\right)^{-1/2} exp\left(\frac{1}{2}\boldsymbol{A}_j^{*T}(\boldsymbol{X}_j^T\boldsymbol{X}_j/\sigma^2 + \boldsymbol{\Sigma}_j^{-1})^{-1}\boldsymbol{A}_j^* - \frac{1}{2}\boldsymbol{\mu}_{\beta_j}^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_{\beta_j}\right)$$

$$\boldsymbol{A}_j^* = (\boldsymbol{X}_j^T\boldsymbol{X}_j\hat{\boldsymbol{\beta}}_j/\sigma^2 + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_{\beta_j})$$

where O'_j is the Bayes factor to include the j term, $\hat{\boldsymbol{\beta}}_j$ are the maximum likelihood estimates of the parameters of the j term.

For unknown σ^2, an alternative prior specification is given by $\boldsymbol{\beta}_j \sim N_{d_j}(0, c^2(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\sigma^2)$ and $\sigma^{-2} \sim G(a_\tau, b_\tau)$ resulting to the Gibbs sampler steps

$$\gamma_j|\sigma^2, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y} \sim Bernoulli\left(\frac{O_j}{1+O_j}\right)$$

$$O_j = \frac{f(\gamma_j = 1, \gamma_{\setminus j})}{f(\gamma_j = 0, \gamma_{\setminus j})}(c^2+1)^{-d_j/2} exp\left(\frac{1}{2\sigma^2}\frac{c^2}{c^2+1}F_j\right)$$

and

$$\sigma^{-2}|\boldsymbol{\gamma}, \boldsymbol{y} \sim G\left(a_\tau + n/2, b_\tau + (\boldsymbol{y}^T\boldsymbol{y} - \frac{c^2}{c^2+1}\sum_{i=1}^{p}\gamma_j F_j)/2\right).$$

If instead of sampling σ^2 we integrate it out we result in Clyde *et al.* (1996) sampler for orthogonal data as given by (3.36).

The main advantage of Clyde (1999) is the computational speed since it is much faster than other MCMC methods. Moreover, it is suitable for ANOVA models where the design matrix is orthogonal. On the other hand, the assumptions of orthogonality and known residual variance required in normal models is not generally the case. The required orthogonality is restrictive and cannot be used when interpretation of casual relationships or selection of a parsimonious model is the main interest. Generally, Smith and Kohn (1996) sampler or MC^3 in linear models can be easily implemented and can handle non-orthogonal data and unknown residual variance.

The extension of Clyde (1999) to non-normal or non-orthogonal models is problematic. The method does not provide good approximations when regressors with low correlation are used. Moreover it does not generally work in Poisson or binomial models since the assumption of constant variance results in bad approximations. A tool should be developed for identifying cases where this method may be applied. An alternative Gibbs sampler for regression models was introduced by Geweke (1996) which was constructed by integrating out from the model selection step for the j term only the corresponding parameter β_j. Summary of the above paragraphs is given as a discussion of Clyde (1999) paper; see Ntzoufras (1999a).

George and McCulloch (1997) developed a sampler similar to Smith and Kohn (1996) sampler based on SSVS. They used a multivariate normal prior for the parameter vector on the full model given by

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma} \sim N(\mathbf{0}_p, [\boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}}]\sigma^2)$$

and the usual gamma prior (3.6) for the residual precision parameter. The resulting posterior is given by

$$\frac{f(\gamma_j = 1|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = (c^2+1)^{-d_j/2} \left(\frac{SS^*_{\gamma_j=1,\boldsymbol{\gamma}_{\setminus j}} + 2b_\tau}{SS^*_{\gamma_j=0,\boldsymbol{\gamma}_{\setminus j}} + 2b_\tau} \right)^{-n/2-a_\tau}, \tag{3.37}$$

where $SS^*_{\boldsymbol{\gamma}}$ are the posterior sum of squares for SSVS given by

$$SS^*_{\boldsymbol{\gamma}} = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + [\boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}}]^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

where $\boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}}$ is the SSVS prior variance defined in (3.28). We may use the simplified prior (3.23) by setting $[\boldsymbol{D}_{\boldsymbol{\gamma}} \boldsymbol{R}_{\boldsymbol{\gamma}} \boldsymbol{D}_{\boldsymbol{\gamma}}]^{-1} = diag[k_j^{2(1-\gamma_j)}\boldsymbol{\Sigma}_j^{-1}]$. In orthogonal cases, using this prior setup, results to a Gibbs sampler similar to Clyde *et al.* (1996) sampler with

$$SS^*_{\boldsymbol{\gamma}} = \boldsymbol{y}^T\boldsymbol{y} - \sum_{j\in\mathcal{V}} k_j^{-2(1-\gamma_j)}\boldsymbol{y}^T\boldsymbol{X}_j\left(\boldsymbol{X}_j^T\boldsymbol{X}_j + \boldsymbol{\Sigma}_j^{-1}\right)^{-1}\boldsymbol{X}_j^T\boldsymbol{y}.$$

In order to make the above equation comparable to (3.36) we use the equivalent to Smith and Kohn (1996) prior that is $\boldsymbol{\Sigma}_j = c^2 \left(\boldsymbol{X}_j^T \boldsymbol{X}_j \right)^{-1}$ resulting to

$$SS^*_{\boldsymbol{\gamma}} = \boldsymbol{y}^T \boldsymbol{y} - \frac{c^2}{c^2+1} \sum_{j \in \mathcal{V}} k_j^{-2(1-\gamma_j)} F_j$$

where F_j is given by (3.35). In this case $SS^*_{\boldsymbol{\gamma}} = SS_{\boldsymbol{\gamma}} + \sum_{j:\gamma_j=0} k_j^{-2} F_j$ and for large k_j we have $SS^*_{\boldsymbol{\gamma}} \approx SS_{\boldsymbol{\gamma}}$. The above fast version of SSVS for orthogonal cases using Smith and Kohn(1996) type prior differs only in the summation where γ_j is substituted by $k_j^{-2(1-\gamma_j)}$. The resulting conditional posterior is given by

$$\frac{f(\gamma_j = 1 | \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0 | \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = (c^2+1)^{-d_j/2} \left(1 - \frac{k_j^2 - 1}{k_j^2} \frac{c^2}{c^2+1} F_j / SS^*_{\gamma_j=0, \boldsymbol{\gamma}_{\setminus j}} \right)^{-n/2}. \quad (3.38)$$

Brown *et al.* (1998) expanded the idea of fast variable selection using SSVS setup in multivariate normal regression models. The resulting posterior for variable selection has similar form to the above involving more complicated matrices. Similar approach can be used to extend Smith and Kohn (1996) methodology to multivariate regression models.

3.5.2 Transformations

A very frequent problem, especially in normal models, is the identification of an appropriate transformation of the response variable Y. Hoeting *et al.* (1995) have used MC^3 and tried to identify which transformation was appropriate. We consider the Box-Cox transformations and therefore the normal linear model is now written

$$Y(\rho) \sim N(\boldsymbol{\eta}, \boldsymbol{I}\sigma^2), \quad Y(\rho) = \begin{cases} (Y^\rho - 1)/\rho & \rho \neq 0 \\ log(Y) & \rho = 0 \end{cases}.$$

Hoeting *et al.* (1995) consider a limited set for values of $\rho \in \{-1, 0, 1/2, 1\}$ but also continuous values can be considered. In the first step we propose to move from m (or $\boldsymbol{\gamma}$) to a new model $m' \in nb(m)$ (or $\boldsymbol{\gamma}'$) which differs by only one covariate (say in j term). Therefore the MC^3 step is given by

$$\alpha = min \left(1, \frac{f(1-\gamma_j | \rho, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y}) j(\boldsymbol{\gamma}', \boldsymbol{\gamma})}{f(\gamma_j, \boldsymbol{\gamma}_{\setminus j} | \rho, \boldsymbol{y}) j(\boldsymbol{\gamma}, \boldsymbol{\gamma}')} \right)$$

while in the second step we propose ρ to change to ρ' with probability $j(\rho', \rho)$ and accept the move with probability

$$\alpha = min \left(1, \frac{f(\rho' | \boldsymbol{\gamma}, \boldsymbol{y}) j(\rho', \rho) j(\rho, \rho')}{f(\rho' | \boldsymbol{\gamma}, \boldsymbol{y}) j(\rho, \rho') j(\rho', \rho)} \right).$$

Alternatively, a Gibbs sampler is given by sampling γ_j from the Bernoulli with success probability $O_j/(1+O_j)$ with O_j equal to (3.31) in which we substitute $SS_{\boldsymbol{\gamma}}$ by $SS_{\rho,\boldsymbol{\gamma}}$ given by

$$SS_{\rho,\boldsymbol{\gamma}} = \boldsymbol{y}(\rho)^T\boldsymbol{y}(\rho) + \boldsymbol{\mu}^T_{\beta_{(\gamma)}}\boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}} - (\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{y}(\rho) + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}})^T\tilde{\boldsymbol{\Sigma}}_{(\gamma)}(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{y}(\rho) + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}}).$$

and

$$f(\rho|\boldsymbol{\gamma},\boldsymbol{y}) \propto [SS_{\rho,\boldsymbol{\gamma}} + 2b_\tau]^{-n/2-a_\tau} f(\rho).$$

For the simplified prior of Smith and Kohn (1996) the posterior residual sum of squares reduces to

$$SS_{\rho,\boldsymbol{\gamma}} = \boldsymbol{y}(\rho)^T\boldsymbol{y}(\rho) - \frac{c^2}{c^2+1}\boldsymbol{y}(\rho)^T\boldsymbol{X}_{(\gamma)}\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{X}_{(\gamma)}\right)^{-1}\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{y}(\rho).$$

The uniform prior on the set of possible values of ρ, $\mathcal{R}$ can be used without any problem. Possible proposal distributions for ρ can be a normal distribution with mean value equal to ρ'.

3.5.3 Outlier Identification

The most common method for outlier identification is called variance inflation method and was used by Hoeting *et al.* (1996) for linear models and Albert and Chib (1997) for generalised linear models. An alternative method is proposed in next chapter. Similar to variable selection procedures, we introduce a latent vector of binary variables $\boldsymbol{v}$ which indicates outliers by $v_i = 0$.

In the variance inflation method, the normal linear model is modified to

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(\gamma)}\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{Q}_{\boldsymbol{v}}\sigma^2\right),$$

where $\boldsymbol{Q}_{\boldsymbol{v}} = Diag[K^{2(1-v_i)}]$ and K is a fixed parameter to be specified. Hoeting *et al.* (1996) suggested $K = 7$ and $f(v_i = 0) = 0.10$ for small datasets ($n < 50$) and $f(v_i = 0) = 0.02$ for larger datasets. They used MC^3 for variable and outlier identification but also Smith and Kohn (1996) approach can be adopted.

Using MC^3 for both variable and outlier identification results in two Metropolis steps. In the first we propose a new model with covariates given by $\boldsymbol{\gamma}'$ differing by $\boldsymbol{\gamma}$ only in one term and accept the move with probability

$$\alpha = min\left(1, \frac{f(1-\gamma_j|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{v}, \boldsymbol{y}) j(\gamma', \gamma)}{f(\gamma_j|\boldsymbol{\gamma}_{\setminus j}, \boldsymbol{v}, \boldsymbol{y}) j(\gamma, \gamma')}\right). \tag{3.39}$$

Then we propose with probability $j(\boldsymbol{v}, \boldsymbol{v}')$ to move from $\boldsymbol{v}$ to $\boldsymbol{v}'$ that differ only in the ith coordinator and accept the move with probability

$$\alpha = min\left(1, \frac{f(1-v_i|\boldsymbol{\gamma}, \boldsymbol{v}_{\setminus i}, \boldsymbol{y}) j(\boldsymbol{v}', \boldsymbol{v})}{f(v_i|\boldsymbol{\gamma}, \boldsymbol{v}_{\setminus i}, \boldsymbol{y}) j(\boldsymbol{v}, \boldsymbol{v}')}\right) \tag{3.40}$$

If a prior of the form

$$\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{v}, \boldsymbol{\gamma} \sim N(\boldsymbol{\mu}_{\beta_{(\gamma)}}, \boldsymbol{\Sigma}_{(\boldsymbol{\gamma}, \boldsymbol{v})})$$

is adopted then the Gibbs sampler is given by sequential generations of γ_j from the Bernoulli with success probability $O_j/(1+O_j)$ with O_j given by (3.31) but the posterior covariance matrix is now given by

$$\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma}, \boldsymbol{v})} = \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{X}_{(\boldsymbol{\gamma})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma}, \boldsymbol{v})}\right)^{-1}$$

while the posterior residual sum of squares $SS_{\boldsymbol{\gamma}}$ is substituted by

$$SS_{\boldsymbol{\gamma}, \boldsymbol{v}} = \boldsymbol{y}^T \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{y} + \boldsymbol{\mu}^T_{\beta_{(\gamma)}} \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma}, \boldsymbol{v})} \boldsymbol{\mu}_{\beta_{(\gamma)}} - (\boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma}, \boldsymbol{v})} \boldsymbol{\mu}_{\beta_{(\gamma)}})^T \tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma}, \boldsymbol{v})} (\boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma}, \boldsymbol{v})} \boldsymbol{\mu}_{\beta_{(\gamma)}}). \tag{3.41}$$

Similarly the outlier identification step will involve sequential generations from similar Bernoulli steps with success probability $O^*_i/(1+O^*_i)$; where O^*_i is given by

$$\begin{aligned} O^*_i &= \frac{f(v_i = 1|\boldsymbol{\gamma}, \boldsymbol{v}_{\setminus i}, \boldsymbol{y})}{f(v_i = 0|\boldsymbol{\gamma}, \boldsymbol{v}_{\setminus i}, \boldsymbol{y})} \\ &= \left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma}, v_i=1, \boldsymbol{v}_{\setminus i})}||\boldsymbol{\Sigma}_{(\boldsymbol{\gamma}, v_i=0, \boldsymbol{v}_{\setminus i})}|}{|\boldsymbol{\Sigma}_{(\boldsymbol{\gamma}, v_i=1, \boldsymbol{v}_{\setminus i})}||\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma}, v_i=0, \boldsymbol{v}_{\setminus i})}|}\right)^{1/2} \left(\frac{SS_{(\boldsymbol{\gamma}, v_i=1, \boldsymbol{v}_{\setminus i})} + 2b_\tau}{SS_{(\boldsymbol{\gamma}, v_i=0, \boldsymbol{v}_{\setminus i})} + 2b_\tau}\right)^{-n/2-a_\tau} \frac{f(v_i = 1, \boldsymbol{v}_{\setminus i})}{f(v_i = 0, \boldsymbol{v}_{\setminus i})}. \end{aligned}$$

If we adopt the prior

$$\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\sigma^2, \boldsymbol{v}, \boldsymbol{\gamma} \sim N\left(\boldsymbol{0}, c^2 \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1} \sigma^2\right)$$

the sampler is reduced to

$$\frac{f(\gamma_j = 1|\boldsymbol{v}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{v}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = (c^2+1)^{-d_j/2} \left(\frac{SS_{\gamma_j=1, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{v}} + 2b_\tau}{SS_{\gamma_j=0, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{v}} + 2b_\tau}\right)^{-n/2-a_\tau} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}$$

and

$$\frac{f(\boldsymbol{y}|v_i = 1, \boldsymbol{v}_{\setminus i}, \gamma)}{f(\boldsymbol{y}|v_i = 0, \boldsymbol{v}_{\setminus i}, \gamma)} = \left(\frac{SS_{v_i=1, \boldsymbol{v}_{\setminus i}, \gamma} + 2b_\tau}{SS_{v_i=0, \boldsymbol{v}_{\setminus i}, \gamma} + 2b_\tau}\right)^{-n/2-a_\tau} \frac{f(v_i = 1, \boldsymbol{v}_{\setminus i})}{f(v_i = 0, \boldsymbol{v}_{\setminus i})}$$

$$SS_{\boldsymbol{v}, \gamma} = \boldsymbol{y}^T \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{y} - \frac{c^2}{c^2+1} \boldsymbol{y}^T \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{X}_{(\boldsymbol{\gamma})} \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1} \boldsymbol{X}^T_{(\boldsymbol{\gamma})} \boldsymbol{Q}^{-1}_{\boldsymbol{v}} \boldsymbol{y}.$$

Chapter 4

Further Developments of MCMC Model and Variable Selection

In this chapter we introduce new MCMC model selection algorithms, describe associations and connections between MCMC methods, develop SSVS priors for factors with multiple categories and log-linear models and provide implementational details. We focus on the methods of Green ('reversible jump', 1995) and Carlin and Chib (1995), and describe a connection between them. We also consider 'variable selection' problems where the models under consideration can be naturally represented by a set of binary indicator variables so that $\mathcal{M} \subseteq \{0,1\}^p$, where p is the total possible number of variables. We introduce a modification of Carlin and Chib's method for variable selection problems, which is more efficient in certain examples. Elements of this chapter and comparisons of MCMC model selection methods are also summarised in two research papers; see Dellaportas *et al.* (1998, 1999).

4.1 Further Gibbs Samplers for Variable Selection

4.1.1 Gibbs Variable Selection

The first approach is a natural hybrid of SSVS and the 'Unconditional Priors' approach of Kuo and Mallick (1998). The linear predictor is assumed to be of the form of (3.2) where, unlike SSVS, variables corresponding to $\gamma_j = 0$ are genuinely excluded from the model. Furthermore, it does not require unnecessary generation from pseudopriors.

We specify the prior for $(\boldsymbol{\beta},\boldsymbol{\gamma})$ as $f(\boldsymbol{\beta},\boldsymbol{\gamma})=f(\boldsymbol{\beta}|\boldsymbol{\gamma})f(\boldsymbol{\gamma})$. If we consider a partition of $\boldsymbol{\beta}$ into $\left\{\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}\right\}$ corresponding to those components of $\boldsymbol{\beta}$ which are included ($\gamma_j=1$) or not included ($\gamma_j=0$) in the model, then the prior $f(\boldsymbol{\beta}|\boldsymbol{\gamma})$ may be partitioned into model prior $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma})$ and pseudoprior $f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\boldsymbol{\gamma})$.

The full conditional posterior distributions are given by

$$f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})},\gamma,\boldsymbol{y}) \;\propto\; f(\boldsymbol{y}|\boldsymbol{\beta},\gamma)f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\gamma)f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\gamma) \tag{4.1}$$

$$f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\gamma,\boldsymbol{y}) \;\propto\; f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\gamma) \tag{4.2}$$

and

$$O_j=\frac{f(\gamma_j=1|\gamma_{\backslash j},\boldsymbol{\beta},\boldsymbol{y})}{f(\gamma_j=0|\gamma_{\backslash j},\boldsymbol{\beta},\boldsymbol{y})}=\frac{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=1,\gamma_{\backslash j})}{f(\boldsymbol{y}|\boldsymbol{\beta},\gamma_j=0,\gamma_{\backslash j})}\,\frac{f(\boldsymbol{\beta}|\gamma_j=1,\gamma_{\backslash j})}{f(\boldsymbol{\beta}|\gamma_j=0,\gamma_{\backslash j})}\,\frac{f(\gamma_j=1,\gamma_{\backslash j})}{f(\gamma_j=0,\gamma_{\backslash j})} \tag{4.3}$$

Note that (4.1) seems less natural than (3.20) as $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})},\gamma,\boldsymbol{y})$ may depend on $\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}$. One way of avoiding this is to assume prior conditional independence of $\boldsymbol{\beta}_j$ terms given $\boldsymbol{\gamma}$, in which case $f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\gamma)$ vanishes from (4.1). This is a restrictive assumption but may be realistic when priors are intended to be non-informative, particularly if the columns of different $\boldsymbol{X}_j$ in (3.2) are orthogonal to each other. Then, each prior for $\boldsymbol{\beta}_j|\boldsymbol{\gamma}$ consists of a mixture of two densities. The first, $f(\boldsymbol{\beta}_j|\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})$, is the true prior for the parameter whereas the second, $f(\boldsymbol{\beta}_j|\gamma_j=0,\boldsymbol{\gamma}_{\backslash j})$, is a pseudoprior. Another way to make (4.1) usable is to define priors $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})},\gamma)=f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\gamma)$ so that we need to calculate the true prior conditional density as $f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\backslash j},\gamma_j=1,\boldsymbol{\gamma}_{\backslash j})$.

This approach is simplified if we assume that the prior for $\boldsymbol{\beta}_j$ depends only on γ_j and is given by

$$f(\boldsymbol{\beta}_j|\gamma_j)=\gamma_j N(\mathbf{0},\boldsymbol{\Sigma}_j)+(1-\gamma_j)N(\bar{\boldsymbol{\mu}}_j,\boldsymbol{S}_j), \tag{4.4}$$

where $\bar{\boldsymbol{\mu}}_j$ and $\boldsymbol{S}_j$ are pseudoprior parameters that can be specified carefully in order to achieve optimal convergence of the MCMC algorithm. The above prior, $f(\boldsymbol{\beta}_j|\gamma_j)$, potentially makes the method less efficient and is most appropriate in examples where $\boldsymbol{X}$ is orthogonal. If prediction, rather than inference about the variables themselves, is of primary interest, then $\boldsymbol{X}$ may always be chosen to be orthogonal (see Clyde *et al.* , 1996).

There is a similarity between this prior and the prior used in SSVS. However, here the full conditional posterior distribution is given by

$$f(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{\setminus j}, \gamma, \boldsymbol{y}) \propto \begin{cases} f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma) N(0, \boldsymbol{\Sigma}_j) & \gamma_j = 1 \\ N(\bar{\boldsymbol{\mu}}_j, \boldsymbol{S}_j) & \gamma_j = 0 \end{cases}$$

and a clear difference between this and SSVS is that the pseudoprior $f(\boldsymbol{\beta}_j|\gamma_j = 0)$ does not affect the posterior distribution and may be chosen as a 'linking density' to increase the efficiency of the sampler, in the same way as the pseudopriors of Carlin and Chib's method. Possible choices of $\bar{\boldsymbol{\mu}}_j$ and $\boldsymbol{S}_j$ may be obtained from a pilot run of the full model; see, for example, Dellaportas and Forster (1999). For more details on selection of pseudoprior parameters see Section 4.5.1.

4.1.2 Variable Selection Using Carlin and Chib Sampler

Here we illustrate how Carlin and Chib (1995) sampler can be simplified for variable selection problems. This variant will be called Carlin and Chib variable selection method (CCVS). We substitute the model indicator m by the term indicator vector $\boldsymbol{\gamma}$ and therefore for model $\boldsymbol{\gamma}$ we have to consider the corresponding parameter vector $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$. This results in

$$O_j^* = \frac{f(\gamma_j = 1|\{\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}, \boldsymbol{\gamma}^* \in \mathcal{M}\}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\{\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}, \boldsymbol{\gamma}^* \in \mathcal{M}\}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = LR_j \times PR_j \times PSR_j \times \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})} \qquad (4.5)$$

where $\{\boldsymbol{\beta}_{\boldsymbol{\gamma}^*}, \boldsymbol{\gamma}^* \in \mathcal{M}\}$ denotes all possible parameter vectors, LR_j, PR_j and PSR_j are the likelihood ratio, the prior and pseudoprior density ratios given by

$$\begin{aligned} LR_j &= \frac{f(\boldsymbol{y}|\boldsymbol{\beta}_{(\gamma_j=1, \boldsymbol{\gamma}_{\setminus j})}, \gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(\gamma_j=0, \boldsymbol{\gamma}_{\setminus j})}, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}, \\ PR_j &= \frac{f(\boldsymbol{\beta}_{(\gamma_j=1, \boldsymbol{\gamma}_{\setminus j})}|\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}_{(\gamma_j=0, \boldsymbol{\gamma}_{\setminus j})}|\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}, \\ PSR_j &= \frac{f(\boldsymbol{\beta}_{(\gamma_j=0, \boldsymbol{\gamma}_{\setminus j})}|\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\boldsymbol{\beta}_{(\gamma_j=1, \boldsymbol{\gamma}_{\setminus j})}|\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})}. \end{aligned}$$

The sampling procedure of Carlin and Chib variable selection sampler can be summarized by

- Generate $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ from the corresponding conditional posterior $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma}, \boldsymbol{y})$.
- For all $j \in \mathcal{V}$ repeat the steps

- Generate $\boldsymbol{\beta}_{(1-\gamma_j,\boldsymbol{\gamma}_{\setminus j})}$ from the pseudoprior $f\left(\boldsymbol{\beta}_{(1-\gamma_j,\boldsymbol{\gamma}_{\setminus j})}|\boldsymbol{\gamma}\right)$.
- Generate γ_j according to (4.5).

The above modification simplifies the Carlin and Chib sampler and makes it efficient for variable selection problems. Now we do not need to generate 2^p vectors from pseudopriors but only the ones necessary for the calculation of (4.5) reducing the number of pseudoprior generations to p. Table 4.1 gives a brief comparison of Gibbs methods. Another approach can be adopted is the Bedrick *et al.* (1996) prior setup as described in Section 3.2.1.5.

4.2 Extensions of Fast Variable Selection Algorithms

4.2.1 Extension to Error Dependent and Autoregressive Models

We can easily extend fast variable selection methodologies (including MC^3) in error dependent models. In the case where the model likelihood can be written as $\boldsymbol{y} \sim N(\boldsymbol{\eta}, \boldsymbol{T}\sigma^2)$ where $\boldsymbol{T}$ denotes a known covariance structure. If we use the MC^3 algorithm then we propose to move from model $\boldsymbol{\gamma}$ to model $\boldsymbol{\gamma}'$ that differ only in j term with probability $j(\boldsymbol{\gamma},\boldsymbol{\gamma}')$ and accept the proposed move with probability equal to

$$\alpha = min\left(1, \frac{f(1-\gamma_j|\boldsymbol{T},\gamma_{\setminus j},\boldsymbol{y})j(\boldsymbol{\gamma}',\boldsymbol{\gamma})}{f(\gamma_j|\boldsymbol{T},\gamma_{\setminus j},\boldsymbol{y})j(\boldsymbol{\gamma},\boldsymbol{\gamma}')}\right).$$

Alternatively, a Gibbs sampler with prior setup given by normal inverse gamma distribution (equations 3.6 and 3.7) will result to sequential generations of γ_j from the Bernoulli with success probability $O_j/(1+O_j)$ with O_j given by (3.31) but the posterior covariance matrix is now given by

$$\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})} = \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{T}^{-1}\boldsymbol{X}_{(\boldsymbol{\gamma})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\right)^{-1}$$

while the posterior residual sum of squares $SS_{\boldsymbol{\gamma}}$ are substituted by

$$SS_{T,\boldsymbol{\gamma}} = \boldsymbol{y}^T\boldsymbol{T}^{-1}\boldsymbol{y} + \boldsymbol{\mu}^T_{\beta_{(\gamma)}}\boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}} - \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{T}^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}}\right)^T \tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})}\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{T}^{-1}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\gamma)}}\right). \tag{4.6}$$

If we use the prior mean $\boldsymbol{\mu}_{\beta_{(\gamma)}} = \mathbf{0}$ and the prior covariance matrix is given by

$$\boldsymbol{\Sigma}_{(\boldsymbol{\gamma})} = c^2\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{T}^{-1}\boldsymbol{X}_{(\boldsymbol{\gamma})}\right)^{-1}\sigma^2$$

then the full conditional posterior odds of j term, O_j, are simplified to

$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{T}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{T}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})} = (c^2+1)^{-d_j/2} \left(\frac{SS_{T,\gamma_j=1,\boldsymbol{\gamma}_{\setminus j}} + 2b_\tau}{SS_{T,\gamma_j=0,\boldsymbol{\gamma}_{\setminus j}} + 2b_\tau} \right)^{-n/2-a_\tau} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})},$$

and the posterior sum of squares is reduced to

$$SS_{T,\boldsymbol{\gamma}} = \boldsymbol{y}^T \boldsymbol{T}^{-1} \boldsymbol{y} - \frac{c^2}{c^2+1} \boldsymbol{y}^T \boldsymbol{T}^{-1} \boldsymbol{X}_{(\gamma)} \left(\boldsymbol{X}_{(\gamma)}^T \boldsymbol{T}^{-1} \boldsymbol{X}_{(\gamma)} \right)^{-1} \boldsymbol{X}_{(\gamma)}^T \boldsymbol{T}^{-1} \boldsymbol{y}.$$

In the case where $\boldsymbol{T}$ is unknown or it has special structure, the model might be given by $\boldsymbol{y} \sim N(\boldsymbol{\eta}, \boldsymbol{T})$ and using the prior $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{T} \sim N\left(\boldsymbol{0}, c^2 \left(\boldsymbol{X}_{(\boldsymbol{\gamma})}^T \boldsymbol{T}^{-1} \boldsymbol{X}_{(\boldsymbol{\gamma})} \right)^{-1} \right)$, results in the posterior distribution

$$f(\boldsymbol{T}, \boldsymbol{\gamma}|\boldsymbol{y}) \propto |\boldsymbol{T}|^{-1/2} exp\left(-\frac{1}{2} SS_{T,\boldsymbol{\gamma}} \right) f(\boldsymbol{T}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}).$$

Fast variable selection methods have been also developed for autoregressive models by Troughton and Godsill (1998) using reversible jump and MC^3 while Barnett *et al.* (1996) used a Gibbs sampler. Additionally, Smith *et al.* (1998) used similar Gibbs based approach in nonparametric regression with autoregressive errors.

4.2.2 Fast Variable Selection Methods for Probit Models

Here we consider the probit regression model for categorical outcomes using continuous latent variables as defined by Albert and Chib (1993). For dichotomous responses Y_i we define the latent values Z_i that satisfy the condition $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ if $Z_i \leq 0$ and assume that Z_i follow a distributional law. If the assumed distribution is normal then we have the probit model $\boldsymbol{Z} \sim N(\boldsymbol{\eta}, \boldsymbol{I})$. In this model formulation we can adopt the prior distribution

$$\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma} \sim N\left(\boldsymbol{0}, c^2 \left(\boldsymbol{X}_{(\boldsymbol{\gamma})}^T \boldsymbol{X}_{(\boldsymbol{\gamma})} \right)^{-1} \right)$$

which results in

$$f(\boldsymbol{Z}, \boldsymbol{\gamma}|\boldsymbol{y}) \propto (c^2+1)^{-d(m)/2} exp\left(-\frac{1}{2} SS_{Z,\boldsymbol{\gamma}} \right) f(\boldsymbol{\gamma}) \prod_{i=1}^{n} \mathcal{E}_i$$

with $\mathcal{E}_i = I(Z_i > 0) I(Y_i = 1) + I(Z_i \leq 0) I(Y_i = 0)$ and $SS_{Z,\boldsymbol{\gamma}}$ are the posterior sum of squares as defined in (3.34) with response values given by the latent vector $\boldsymbol{Z}$. The Gibbs sampler can be constructed by two steps

$$f(\boldsymbol{Z}|\boldsymbol{\gamma}, \boldsymbol{y}) = N_n \left(\boldsymbol{0}, \left[\boldsymbol{I} - \frac{c^2}{c^2+1} \boldsymbol{X}_{(\boldsymbol{\gamma})} \left(\boldsymbol{X}_{(\boldsymbol{\gamma})}^T \boldsymbol{X}_{(\boldsymbol{\gamma})} \right)^{-1} \boldsymbol{X}_{(\boldsymbol{\gamma})}^T \right]^{-1} \right) \prod_{i=1}^{n} \mathcal{E}_i$$

$$\frac{f(\gamma_j=1|\boldsymbol{\gamma}_{\setminus j},\boldsymbol{Z},\boldsymbol{y})}{f(\gamma_j=0|\boldsymbol{\gamma}_{\setminus j},\boldsymbol{Z},\boldsymbol{y})} = (c^2+1)^{-d_j/2} exp\left(-\frac{1}{2}[SS_{Z,\gamma_j=1,\boldsymbol{\gamma}_{\setminus j}} - SS_{Z,\gamma_j=0,\boldsymbol{\gamma}_{\setminus j}}]\right) f(\boldsymbol{\gamma}).$$

For polychotomous responses the procedure is similar but the limits $\mathcal{E}_i$ should be adjusted equivalently. When other distributions, such as Student are preferred, then a Gibbs variable or a reversible jump setup should be adopted. Similar approaches can be adopted if we want to incorporate distribution selection or other characteristics.

4.3 Connections Between Markov Chain Monte Carlo Model Selection Methods

Special cases of reversible jump sampler are described in detail here. The general sampling scheme of the algorithm is given in Section 3.3.1 in which we can substitute m by $\boldsymbol{\gamma}$ when we are interested in covariate selection only. Metropolised Carlin and Chib method is introduced as a special case of reversible jump. Work on the relationships of MCMC methods was also reported by Godsill (1998).

4.3.1 Reversible Jump and 'Metropolised' Carlin and Chib

The Gibbs sampler proposed by Carlin and Chib (1995) requires the calculation of all A_m in the denominator of (3.21). An alternative approach is a hybrid Gibbs/Metropolis strategy, where the 'model selection' step is not based on the full conditional, but on a proposal for a move to model m', followed by acceptance or rejection of this proposal. If the current state is model m and we propose model m' with probability $j(m,m')$, then the acceptance probability is given by

$$\begin{aligned} \alpha &= \min\left(1, \frac{A_{m'}j(m',m)}{A_m j(m,m')}\right) \\ &= \min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}_{(m')},m')f(\boldsymbol{\beta}_{(m')}|m')f(\boldsymbol{\beta}_{(m)}|m')f(m')j(m',m)}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)},m)f(\boldsymbol{\beta}_{(m)}|m)f(\boldsymbol{\beta}_{(m')}|m)f(m)j(m,m')}\right) \end{aligned} \quad (4.7)$$

as all other pseudopriors cancel.

Note that when we are in model m and we propose model m', we require only values of $\boldsymbol{\beta}_{(m)}$ and $\boldsymbol{\beta}_{(m')}$ to calculate α in (4.7). Furthermore, we are assuming that model m' is proposed with probability $j(m,m')$, independently of the values of any model parameters.

Therefore if we reverse the order of sampling from $j(m, m')$ and the full conditional distributions for $\boldsymbol{\beta}_{(m_k)}$ in (3.20), there is no need to sample from any pseudopriors other than that for m'. The method now consists of the following three steps

- Propose a new model m' with probability $j(m, m')$.
- Generate $\boldsymbol{\beta}_{(m')}$ from the pseudoprior $f(\boldsymbol{\beta}_{(m')}|m \neq m')$.
- Accept the proposed move to model m' with probability α given by (4.7).

It is straightforward to see that by a simple modification ('Metropolising' the model selection step), Carlin and Chib's method becomes a special case of reversible jump with

$$\left(\boldsymbol{\beta}'_{(m')}, \boldsymbol{u}'\right) = \left(\boldsymbol{u}, \boldsymbol{\beta}_{(m)}\right), \ \boldsymbol{u}' = \{\boldsymbol{\beta}_{(m_l)} : m_l \neq m'\}, \ \boldsymbol{u} = \{\boldsymbol{\beta}_{(m_l)} : m_l \neq m\},$$

where $\{\boldsymbol{\beta}_{(m_l)} : m_l \neq m\}$ are the parameter vectors $\boldsymbol{\beta}_{(m_l)}$ for all $m_l \in \mathcal{M} \setminus \{m\}$, while the proposal densities are replaced by

$$q(\boldsymbol{u}|\boldsymbol{\beta}_{(m)}, m, m') = \prod_{m_l \in \mathcal{M}\setminus\{m'\}} \left\{f(\boldsymbol{\beta}_{(m_l)}|m')\right\}$$

and

$$q(\boldsymbol{u}'|\boldsymbol{\beta}'_{(m')}, m', m) = \prod_{m_l \in \mathcal{M}\setminus\{m\}} \left\{f(\boldsymbol{\beta}_{(m_l)}|m)\right\}.$$

We should further note that the above constructed reversible jump (and the equivalent Metropolised Carlin and Chib) coincide to a simpler reversible jump scheme with $\left(\boldsymbol{\beta}'_{(m')}, \boldsymbol{u}'\right) = \left(\boldsymbol{u}, \boldsymbol{\beta}_{(m)}\right)$, $\boldsymbol{u}' = \boldsymbol{\beta}_{(m)}$, $\boldsymbol{u} = \boldsymbol{\beta}'_{(m')}$, and proposal $q(\boldsymbol{u}|\boldsymbol{\beta}_{(m)}, m, m')$ replaced by pseudoprior $f(\boldsymbol{\beta}_{(m')}|m \neq m')$.

4.3.2 Using Posterior Distributions as Proposals

Suppose that, for each m, the posterior density $f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})$ is available, including the normalising constant which is the marginal likelihood $f(\boldsymbol{y}|m)$. If this distribution is used as a pseudoprior then the acceptance probability in (4.7) is given by

$$\begin{aligned}
\alpha &= min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{(m')}, m')f(\boldsymbol{\beta}'_{(m')}|m')f(m')j(m', m)f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m)f(\boldsymbol{\beta}_{(m)}|m)f(m)j(m, m')f(\boldsymbol{\beta}'_{(m')}|m', \boldsymbol{y})}\right) \\
&= min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{(m')}, m')f(\boldsymbol{\beta}'_{(m')}|m')f(m')j(m', m)f(\boldsymbol{\beta}_{(m)}, m, \boldsymbol{y})/f(m, \boldsymbol{y})}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m)f(\boldsymbol{\beta}_{(m)}|m)f(m)j(m, m')f(\boldsymbol{\beta}'_{(m')}, m', \boldsymbol{y})/f(m', \boldsymbol{y})}\right) \\
&= min\left(1, B_{m'm}\frac{f(m')j(m', m)}{f(m)j(m, m')}\right)
\end{aligned}$$

where $B_{m'm}$ is the Bayes factor of model m' against model m. In practice, we can not usually calculate $B_{m'm}$. In the special case where models are decomposable graphical models, Madigan and York (1995) used exactly this approach, which they called MC^3. From the above it is clear that MC^3 is a special case of both Metropolised Carlin and Chib and reversible jump algorithms. Here there is no need to generate the model parameters $\boldsymbol{\beta}_{(m)}$ as part of the Markov chain. These can be generated separately from the known posterior distributions $f(\boldsymbol{\beta}_{(m)}|m, \boldsymbol{y})$ if required.

4.3.3 Reversible Jump for Covariate Selection

When model selection is restricted in covariate selection and the regressors are not highly correlated then we can use simple reversible jump. In this simple version of reversible jump we may substitute the model indicator m by the corresponding vector of binary indicators $\boldsymbol{\gamma}$. In such case the proposal $j(m, m')$ is substituted by $j(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$. Usually the strategy to move in neighbourhood models as defined by Madigan and York (1995) and was also used by Dellaportas and Forster (1999) is adopted. In such case $j(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$ can be substituted by the product $q_1(\boldsymbol{\gamma}, j) \times q_2(\gamma_j, 1-\gamma_j)$. The first proposal q_1 denotes the probability we propose to change j term when we are in model $\boldsymbol{\gamma}$ while q_2 denotes the probability for this term to change from γ_j to $1-\gamma_j$. Usual choices are $q_1(\boldsymbol{\gamma}, j) = 1/p$ and $q_2(\gamma_j, 1-\gamma_j) = 1.0$, that is, select equal probability one term j and alway propose to change it; for further details see Section 4.5.1.2.

This simplified version of reversible jump is summarized by the following steps

- Generate $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ from the full conditional posterior distribution $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\gamma, \boldsymbol{y})$.
- Select a candidate term j to change with probability $1/p$ and propose to change from γ_j to $1-\gamma_j$ with probability one.
- If $\gamma_j = 0$ then

  [a] Generate the additional parameters $\boldsymbol{\beta}'_j$ from the proposal density $q_j(\boldsymbol{\beta}'_j)$,

  [b] Set $\boldsymbol{\beta}'_{(\gamma')} = [\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{\beta}'_j]$ and

  [c] Accept the proposed move with probability

$$\alpha = min\left(1, \frac{f(\boldsymbol{y}|\gamma', \boldsymbol{\beta}'_{(\boldsymbol{\gamma}')})f(\boldsymbol{\beta}'_{(\boldsymbol{\gamma}')}|\gamma')f(\gamma')}{f(\boldsymbol{y}|\gamma, \boldsymbol{\beta}_{(\boldsymbol{\gamma})})f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\gamma)f(\gamma)q_j(\boldsymbol{\beta}'_j)}\right). \tag{4.8}$$

- If $\gamma_j = 1$ then

  [a] Set $\boldsymbol{\beta}'_{(\gamma')}$ equal to the parameters of $\boldsymbol{\beta}_{(\gamma)}$ removing the parameters that correspond to j term and

  [b] Accept the proposed move with probability

$$\alpha = min\left(1, \frac{f(\boldsymbol{y}|\gamma', \boldsymbol{\beta}'_{(\gamma')})f(\boldsymbol{\beta}'_{(\gamma')}|\gamma')f(\gamma')q_j(\boldsymbol{\beta}_j)}{f(\boldsymbol{y}|\gamma, \boldsymbol{\beta}_{(\gamma)})f(\boldsymbol{\beta}_{(\gamma)}|\gamma)f(\gamma)}\right). \tag{4.9}$$

If a different proposal scheme for the model indicators is desired then must propose the new model with probability to move from model $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}'$ and further multiply the acceptance probabilities by the ratio $j(\boldsymbol{\gamma}', \boldsymbol{\gamma})/j(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$.

4.3.4 Metropolis within Gibbs Variable Selection

Gibbs variable selection as defined in Section 4.1.1 can be substituted by the corresponding Metropolis within Gibbs step. Assuming that γ_j and $\gamma'_j = 1 - \gamma_j$ are the current and the proposed values of the indicator variable corresponding to $\boldsymbol{\beta}_j$, respectively, then the Metropolis acceptance probability is given by

$$\alpha = min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma'_j, \boldsymbol{\gamma}_{\setminus j})f(\boldsymbol{\beta}|\gamma'_j, \boldsymbol{\gamma}_{\setminus j})f(\gamma'_j, \boldsymbol{\gamma}_{\setminus j})j(\gamma'_j, \gamma_j)}{f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma_j, \boldsymbol{\gamma}_{\setminus j})f(\boldsymbol{\beta}|\gamma_j, \boldsymbol{\gamma}_{\setminus j})f(\gamma_j, \boldsymbol{\gamma}_{\setminus j})j(\gamma_j, \gamma'_j)}\right) \tag{4.10}$$

where $j(\gamma_j, \gamma'_j)$ and $j(\gamma'_j, \gamma_j)$ denote the probability of proposing the terms γ'_j and γ_j respectively.

Each of the above Metropolis within Gibbs step is a reversible jump step. Suppose that we propose to move to 'neighbourhood' models (models that differ in one term) and that the current model m corresponds to an indicator vector $\boldsymbol{\gamma}$ and the proposed model m' to the vector $\boldsymbol{\gamma}'$ then

$$\exists \;\; j \in \mathcal{V}: \quad \gamma'_j = 1 - \gamma_j \;\; \text{and} \;\; \gamma'_l = \gamma_l, \quad \forall \;\; l \in \mathcal{V} \setminus \{j\}.$$

From the above we have that

$$f(m) = f(\boldsymbol{\gamma}) = f(\gamma_j, \boldsymbol{\gamma}_{\setminus j}), \quad f(m') = f(\boldsymbol{\gamma}') = f(\gamma'_j, \boldsymbol{\gamma}_{\setminus j}).$$

and

$$j(m, m') = j(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = q_1(j)q_2(\gamma_j, \gamma'_j)$$

where $q_1(j)$ is the probability to propose j term to change and $q_2(\gamma_j, \gamma'_j)$ is the probability that γ_j will change to γ'_j.

We will now show that the acceptance probability given in (4.10) is a special case of the reversible jump with acceptance probability given by (3.19). First, we split the parameter vector $\boldsymbol{\beta}$ into vectors $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ and $\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}$ elements of $\boldsymbol{\beta}$ which are included or not in model m respectively. Thus, $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ contains only the elements $\boldsymbol{\beta}_j$ with $\gamma_j = 1$ while $\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}$ contains the remaining elements (with $\gamma_j = 0$). The above definition implies that $\boldsymbol{\beta}_{(m)} = \boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ and therefore the prior distribution $f(\boldsymbol{\beta}_{(m)}|m)$ is given by the density $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma})$. Using $\boldsymbol{u} = \boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}$ and as proposal the pseudoprior $f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma})$ results in

$$f(\boldsymbol{\beta}_{(m)}|m)q(\boldsymbol{u}|\boldsymbol{\beta}_{(m)}, m', m) = f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma})f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma}) = f(\boldsymbol{\beta}|\boldsymbol{\gamma}) = f(\boldsymbol{\beta}|\gamma_j, \boldsymbol{\gamma}_{\backslash j})$$

and

$$f(\boldsymbol{\beta}'_{(m')}|m')q(\boldsymbol{u}'|\boldsymbol{\beta}'_{(m')}, m, m') = f(\boldsymbol{\beta}_{(\boldsymbol{\gamma}')}|\boldsymbol{\gamma}')f(\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma}')}|\boldsymbol{\beta}_{(\boldsymbol{\gamma}')}, \boldsymbol{\gamma}') = f(\boldsymbol{\beta}|\boldsymbol{\gamma}') = f(\boldsymbol{\beta}|\gamma'_j, \boldsymbol{\gamma}_{\backslash j}).$$

Finally, it is clear that for the likelihood terms in (3.19) we have

$$f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m) = f(\boldsymbol{y}|\boldsymbol{\beta}_{(\gamma)}, \boldsymbol{\gamma}) = f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma_j, \boldsymbol{\gamma}_{\backslash j}),$$

$$f(\boldsymbol{y}|\boldsymbol{\beta}'_{(m')}, m') = f(\boldsymbol{y}|\boldsymbol{\beta}_{(\gamma')}, \boldsymbol{\gamma}') = f(\boldsymbol{y}|\boldsymbol{\beta}, \gamma'_j, \boldsymbol{\gamma}_{\backslash j}).$$

Substituting the above equations in (3.19) the acceptance probability is equal to (4.10). If in Gibbs variable selection we use priors (4.4) and pseudopriors given by

$$f(\boldsymbol{\beta}_j|\gamma_j, \boldsymbol{\gamma}_{\backslash j}) = f(\boldsymbol{\beta}_j|\gamma_j) \tag{4.11}$$

then the acceptance probability (4.10) simplifies to reversible jump one of Section 4.3.3. The number of parameters that we need to generate drops from the dimension of the full model model, to $d(\boldsymbol{\gamma}) + (1-\gamma_j)d_j$ which is the maximum number of parameters between models $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$.

The above Metropolis within Gibbs variable selection step is equivalent to a Metropolised Gibbs sampler as defined by Liu (1996a,b), if we use $q_2(\gamma_j, 1-\gamma_j) = 1$. This version of Metropolis within Gibbs variable selection, according to Liu (1996a,b), is optimal and reaches convergence quicker than any other Metropolis or Gibbs sampler.

4.4 Comparison of Variable Selection Methods

Expression (4.3) is similar to expressions (3.25) and (3.30) in other proposed variable selection methods. George and McCulloch (1993) propose SSVS strategy which assumes the maximal model throughout, but constrains $\boldsymbol{\beta}_j$ parameters to be close to zero when $\gamma_j = 0$. In this situation, $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma})$ is independent of $\boldsymbol{\gamma}$ and so the first ratio on the right hand side of (4.3) vanishes. Kuo and Mallick (1998) propose a similar approach to the above but use a prior distribution for $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with $\boldsymbol{\beta}$ independent of $\boldsymbol{\gamma}$. Then, the second term on the right hand side of (4.3) vanishes. For this reason, Gibbs variable selection can be though as a trade off between SSVS and Kuo and Mallick sampler since the variable selection step depends on both likelihood and on prior densities ratios. The association between the variable selection Gibbs samplers are summarized in Table 4.1.

Carlin and Chib's method involves a single model indicator parameter. Therefore, at each iteration of the Gibbs sampler all parameters $\boldsymbol{\beta}_{(m)}$ of all models $m \in \mathcal{M}$ are generated from either posterior distribution or pseudoprior and the model selection step allows a simultaneous change of all γ_js. In Gibbs variable selection, an observation of $\boldsymbol{\gamma}$ is generated following generation of the whole parameter vector $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\beta}_{(\setminus\boldsymbol{\gamma})}\right]$ from either the posterior distributions for $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ or the pseudoprior densities for $\boldsymbol{\beta}_{(\setminus\boldsymbol{\gamma})}$. This procedure will generally involve generating each term parameter vector $\boldsymbol{\beta}_j$ from p conditional distributions, a much smaller burden than required for Carlin and Chib's method. Furthermore, it would seem to be more efficient to generate pairs of $(\boldsymbol{\beta}_j, \gamma_j)$ successively, possibly by a random scan, so that more local moves in model space are attempted.

The modified version of Carlin and Chib sampler for variable selection is also efficient and faster than the original method. The major difference between Gibbs variable selection and this modified algorithm is that in the former we specify pseudopriors only for the additional terms while in the latter for the whole parameter vector of each model. Although the latter is a drawback in terms of computing time it gives to Carlin and Chib type of variable selection sampler the flexibility to handle problems with highly correlated regressors where the simpler Gibbs variable selection demonstrates convergence difficulties.

Clearly, moves between models m $(\boldsymbol{\gamma})$ and m' $(\boldsymbol{\gamma}')$ may also be based on a Metropolis step, as was suggested in Section 4.3.1. Then the pseudopriors may be thought of as part of the proposal density for parameters which are present in one model but not in the other. This highlights a drawback of the variable selection approaches discussed in this section, namely that parameters which are 'common' to both models remain unchanged, and therefore the procedure will not be efficient unless posterior distributions of such parameters are similar under both models. Note that Gibbs variable selection corresponds to the simple version of reversible jump for variable selection while Carlin and Chib type variable selection sampler corresponds to Metropolised version of Carlin and Chib sampler.

		O_j		
Method	$\boldsymbol{\eta}$	PSR_j	LR_j	PR_j
SSVS	$\boldsymbol{X\beta}$			$\surd$
KM	$\sum \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$		$\surd$	
GVS	$\sum \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$	$\surd$	$\surd$	$\surd$
CCVS	$\boldsymbol{X}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_{(\boldsymbol{\gamma})}$	$\surd$	$\surd$	$\surd$

Table 4.1: Components of Full Conditional Posterior Odds for Inclusion of Term j, O_j, in Each Variable Selection Algorithm (PSR = Pseudoprior Ratio, LR = Likelihood Ration, PR = Prior Density Ratio).

4.5 Further Considerations

This section discusses aspects of special interest involved in MCMC model selection methods including proposed strategies for specification of proposal and pseudoprior densities and parametrizations that we should use.

4.5.1 Proposal Distributions

A crucial aspect in MCMC for most model selection methods described in this thesis is the choice of proposal or pseudoprior distributions which control the convergence rates of the resulted chains. For this reason, we propose simple rules for the selection of proposal or pseudoprior distributions for generalised linear models. Two subsections are presented;

the first suggests proposal distributions for model parameters, while the second for model indicator m.

4.5.1.1 Proposal Distributions for Model Parameters

The simplest way to specify of proposal or pseudoprior distributions is to consider independent normal distributions for each term j given by

$$N(\bar{\boldsymbol{\mu}}_j, \boldsymbol{S}_j) \tag{4.12}$$

with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{S}_j$ estimated from a small pilot run of the full model. In most cases, MCMC using these proposal distributions performs well but may exhibit convergence difficulties in cases where highly correlated regressors are considered.

A second 'automatic' specification of proposal distributions is to consider distributions of type

$$N(\mathbf{0}_{d_j}, \boldsymbol{\Sigma}_j/k_j^2)$$

with covariance equal to the prior covariance matrix divided by the pseudoprior parameter k_j^2 following the notion of SSVS prior setup (3.23). In such case, the parameters proposed by the MCMC algorithm will be within a neighborhood of zero. The parameter k_j is now a pseudoprior parameter which controls the area of the proposed values and large values of k_j slow down the convergence of the chain. The choice of $k_j = 10$ appears a good default choice for the pseudoprior specification which performs sufficiently well in most cases. This approach is closely related to the automatic choices proposed by Giudici and Roberts (1998) in reversible jump sampler.

A better proposal can be based in model specific maximum likelihood estimates and therefore for the normal model we may use

$$q(\boldsymbol{\beta}'_{(m')}|\sigma^2, \boldsymbol{\beta}_{(m)}, m', m) = N\left(\hat{\boldsymbol{\beta}}_{(m)}, \left(\boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}\right)^{-1}\sigma^2\right)$$

where $\hat{\boldsymbol{\beta}}_{(m)}$ are the maximum likelihood estimates under model m. Under the generalised linear model specification we can use as a proposal

$$q(\boldsymbol{\beta}'_{(m')}|\boldsymbol{\beta}_{(m)}, m', m) = N\left(\hat{\boldsymbol{\beta}}_{(m)}, \left(\boldsymbol{X}_{(m)}^T \hat{\boldsymbol{H}}_{(m)} \boldsymbol{X}_{(m)}\right)^{-1}\right)$$

where $\hat{\boldsymbol{H}}_{(m)}$ is given by (3.12). An alternative easy-to-use choice is given by

$$q(\boldsymbol{\beta}'_{(m')}|\boldsymbol{\beta}_{(m)}, m', m) = N(\hat{\boldsymbol{\beta}}_{(m)}, \boldsymbol{\Sigma}_{(m)}/k^2).$$

The above choice greatly simplifies the computations in the model selection step and provides the possibility to control the convergence of the chain via the parameter k^2. These proposals complicate the model selection step since, in each iteration, we need to calculate the maximum likelihood estimates for each proposed model but it highly increases the efficiency of the MCMC algorithm.

In the case of simple reversible jump we may alternatively propose additional terms in such a way that the likelihood of the proposed model m' is maximized, conditionally on the rest model parameters. Therefore we have

$$q(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{(\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}, \gamma_j = 1, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j}) = N\left(\left(\boldsymbol{X}_j^T\hat{\boldsymbol{H}}\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j^T\hat{\boldsymbol{H}}\boldsymbol{\eta}_j^*, \left(\boldsymbol{X}_j^T\hat{\boldsymbol{H}}\boldsymbol{X}_j\right)^{-1}\right),$$

where $\hat{\boldsymbol{H}}$ is the weight matrix used in observed information matrix of the 'saturated' model and $\boldsymbol{\eta}_j^*$ is a vector with elements given by

$$\{\boldsymbol{\eta}_j^*\}_i = g(y_i) - \sum_{l\in\mathcal{V}\setminus\{j\}} \gamma_l \boldsymbol{x}_{il}\boldsymbol{\beta}_l.$$

When the j term is univariate then the above proposal is simply given by

$$q(\beta_j|\boldsymbol{\beta}_{(\gamma_j=0,\boldsymbol{\gamma}_{\setminus j})}, \gamma_j = 1, \gamma_j = 0, \boldsymbol{\gamma}_{\setminus j}) = N\left(\frac{\sum_{i=1}^n x_{ij}h_i\{\boldsymbol{\eta}_j^*\}_i}{\sum_{i=1}^n h_i x_{ij}^2}, \frac{1}{\sum_{i=1}^n h_i x_{ij}^2}\right).$$

Alternatively, for simplicity, we may substitute the covariance matrix by $\boldsymbol{\Sigma}_j/k^2$.

Giudici and Roberts (1998) developed a method for automatic choice of scale parameter in proposal distributions used in reversible jump for nested models. This scale parameter varies in each iteration according to the proposed values and maximizes the acceptance probability when the proposed parameters are considered equal to zero.

4.5.1.2 Proposal Distributions on Model Space

The most common proposal distribution used for model space is the uniform distribution over a restrictive set of models. This restrictive set of models was originally called by Madigan and York (1995) 'neighborhood of model m' including models that differ by one variable. Dellaportas and Forster (1999) also used similar restrictive set of proposed models. This kind of proposal distributions will be called 'local' while proposals that consider all possible models will be called 'global'.

Global model proposals result in low acceptance rates and therefore local proposals are preferred especially there is some natural structure in the model space, for example nested models. Generally, reversible jump chains with local proposals perform well but may exhibit difficulties in some ill-posed problems for example when highly correlated regressors are included in one model. In such cases using a combination of local and global proposals may be an optimal choice.

The choice of the value of the probability $j(m, m)$ is of crucial interest. Liu (1996a,b) suggested an improved Metropolis sampler for discrete random variables which combines the advantages of Gibbs and Metropolis algorithms. According to his work it is more efficient to use a Metropolis sampler in which we propose the same value with probability zero and and the rest of the outcomes with probability equal to the full conditional posterior given that the probability propose the same value is constraint to zero.

We propose to use his arguments in the variable selection samplers used in this thesis. In the variable selection samplers we may adopt the following two sampling steps

1. *Random scan Gibbs variable selection*: Pick a latent term indicator γ_j at random and update it from $f(\gamma_j|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})$.

2. *Simple Reversible Jump for variable selection*: Propose to move to a new model $\boldsymbol{\gamma}' \in nb(\boldsymbol{\gamma})$ (or $m' \in nb(m)$) which is equivalent to pick an indicator term γ_j and propose to change it to $1 - \gamma_j$. Therefore vector $\boldsymbol{\gamma}'$ differs from current $\boldsymbol{\gamma}$ only in the jth coordinate. Accept the move with probability given by (4.8) or (4.9).

If we adopt the ideas of Liu (1996a,b) then a Metropolis algorithm always proposing to change γ_j to $1 - \gamma_j$ is better than the corresponding Gibbs algorithm. Furthermore, we argue that practical work in generalised linear models has indicated that the choice of $j(m, m) = 0$ is more efficient than $j(m, m) > 0$.

For this reason, reversible jump within Gibbs variable selection with $q_2(\gamma_j, 1 - \gamma_j) = 1$ should be preferred. Similar arguments can be used for the variants of Carlin and Chib sampler introduced in this thesis (Metropolised and variable selection versions). Generally, the choice of $j(m,m)=0$ should be preferred. With similar arguments we can use Metropolis steps for updating each γ_j in Smith and Kohn (1996) resulting to an optimal MCMC strategy.

Under this choice the above proposal slightly changes to

$$j(m',m) = 1/|nb(m')|$$

where $nb(m)$ is the set of models that differ from m in one term.

A good choice for model proposals may be obtained by using either Laplace or BIC approximation. For example we may use the proposal

$$j(m',m) = \frac{(2\pi)^{\frac{d(m)}{2}}|\boldsymbol{X}^T_{(m)}\hat{\boldsymbol{H}}_{(m)}\boldsymbol{X}_{(m)}|^{\frac{1}{2}} f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m)},m) f(\hat{\boldsymbol{\beta}}_{(m)}|m)}{\sum\limits_{m_l \in nb(m')} (2\pi)^{\frac{d(m_l)}{2}}|\boldsymbol{X}^T_{(m_l)}\hat{\boldsymbol{H}}_{(m_l)}\boldsymbol{X}_{(m_l)}|^{\frac{1}{2}} f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m_l)},m) f(\hat{\boldsymbol{\beta}}_{(m_l)}|m)}$$

based on the Laplace approximation or

$$j(m',m) = \frac{n^{-d(m)/2}[f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m)},m)]}{\sum\limits_{m_l \in nb(m')} n^{-d(m_l)/2}[f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_{(m_l)},m_l)]}$$

based on BIC approximation. The latter must be handled with care since in some cases may not give accurate approximations.

Similarly in 'global' moves we may use any of the above approximations. In cases where the number of models is huge we may use an MC^3 type algorithm based on BIC or Laplace approximations to get rough estimates of posterior weights. This may increase the computation effort but improve the efficiency of the reversible jump.

4.5.2 Parametrizations and Data Transformations

It is clear that when the data matrix $\boldsymbol{X}_{(m)}$ is orthogonal model selection becomes straightforward; see for implementation Foster and George (1994), Clyde *et al.* (1996) and Clyde (1999). Therefore, if our main interest is prediction rather than interpretation of casual relationships, then orthogonalizing is the ideal solution to our problems. Parameters have similar interpretation across all models and simple methods with straightforward proposals can be used without any difficulty; see Clyde *et al.* (1996) and Clyde (1999). On the other hand, MC^3 and Smith and Kohn (1996) samplers for normal models can handle non-orthogonal data and provide accurate results very fast. Therefore, for normal models there is no need for orthogonalizing. In generalised linear models the problem is more complicated.

Since orthogonalizing simplifies model selection procedures, we shall adopt orthogonal constraints when categorical factors are considered as possible regressors. Such an approach was used by Dellaportas and Forster (1999) for log-linear model selection.

Another crucial question is whether we should standardise all variables. This will result in a new transformed model. Moreover, using priors on the transformed model is straightforward since each model coefficient has similar interpretation (that is, as X_j increases by one standard deviation, s_j, Y will increase by β_j times s_y; where s_y is the standard deviation of Y). This approach was adopted by Raftery *et al.* (1997). Although, standardizing may solve some prior specification problems, it does not solve all problems appearing in model selection since possible correlations between model parameters are not eliminated.

4.6 Implementation of MCMC Variable Selection Algorithms in Generalised Linear Models

A very popular model formulation is given by the generalised linear models. For this reason, a lot of work have been published on Bayesian model selection for members of the generalised linear models including Lindley (1968), Atkinson (1978), Smith and Spiegelhalter (1980), Spiegelhalter and Smith (1982), Mitchell and Beauchamp (1988), Albert (1991, 1996) and Raftery (1996a). MCMC samplers were developed for generalised linear models after the early nineties and include George and McCulloch (1993, 1996, 1997), George *et al.* (1996), Carlin and Chib (1995), Green (1995), Hoeting *et al.* (1995, 1996), Smith and Kohn (1996), Clyde *et al.* (1996), Geweke (1996), Chipman (1996, 1997), Chipman *et al.* (1997), Raftery *et al.* (1997), Clyde and DeSimone-Sasinowska (1997), Clyde (1999), Troughton and Godsill (1997), Albert and Chib (1997), Kuo and Mallick (1998), Petris and Tardella (1998) and Dellaportas and Forster (1999).

The aim of this section is to formulate a general frame under which all MCMC methods can be summarized. Under this framework we can easily understand the peculiarities, usefulness and the working mechanism of each algorithm.

The general form of the likelihood of a generalised linear model is given by

$$f(\boldsymbol{y}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \phi, \boldsymbol{\gamma}) = exp\left\{\sum_{i=1}^{n} \frac{y_i g^*(\eta_i) - b\{g^*(\eta_i)\}}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, a_i(\phi))\right\} \tag{4.13}$$

where η_i is the linear predictor for i observation and is given by different equation depending the method used, $g^*(x)$ is function connecting the parameter ϑ_i of the exponential family and the linear predictor and is given by $\vartheta_i = \vartheta(\mu_i) = \vartheta(g^{-1}(\eta_i))$ [hence $g^*(x) = \vartheta(g^{-1}(x))$] and $g(x)$ is the link function connecting the expected value of y_i with the linear predictor η_i. For the Gibbs variable selection

$$\boldsymbol{\eta} = \sum_{j \in \mathcal{V}} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j \tag{4.14}$$

which is equivalent to

$$\boldsymbol{\eta} = \boldsymbol{X}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_{(\boldsymbol{\gamma})} \tag{4.15}$$

where $\boldsymbol{X}_{(\boldsymbol{\gamma})}$ and $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ denote the design or data matrix and the vector of coefficients constructed from all terms included in the model. The above linear predictor is also used in Carlin and Chib type of samplers but each $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ takes different values. In SSVS the model indicator $\boldsymbol{\gamma}$ is not involved in the linear predictor (or generally in the likelihood) and therefore

$$\boldsymbol{\eta} = \boldsymbol{X} \boldsymbol{\beta} \tag{4.16}$$

for all models. The inclusion of the variables is greatly controlled via the prior distribution used.

To complete the Bayesian formulation of generalised linear models we use prior distributions discussed in Section 3.2. Although we may prefer to adopt independent prior distributions for their simplicity and straightforward interpretation, we generally should avoid them since they affect the posterior model probabilities; see Chapter 6.

Generally the procedure of MCMC methods can be summarized in the following steps:

1. Generate $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ from $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})} | \phi, \boldsymbol{\gamma}, \boldsymbol{y}) \propto$

$$exp \left[\sum_{i=1}^{n} \frac{y_i g^* \left([\boldsymbol{X}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_{(\boldsymbol{\gamma})}]_i \right) - b \left\{ g^* \left([\boldsymbol{X}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_{(\boldsymbol{\gamma})}]_i \right) \right\} }{a_i(\phi)} - \frac{1}{2} \left(\boldsymbol{\beta}_{(\boldsymbol{\gamma})} - \boldsymbol{\mu}_{\beta_{(\gamma)}} \right)^T \boldsymbol{\Sigma}_{(\boldsymbol{\gamma})}^{-1} \left(\boldsymbol{\beta}_{(\boldsymbol{\gamma})} - \boldsymbol{\mu}_{\beta_{(\gamma)}} \right) \right]$$

where $[\boldsymbol{X}_{(\boldsymbol{\gamma})} \boldsymbol{\beta}_{(\boldsymbol{\gamma})}]_i$ is the linear predictor for i observation and is substituted by $[\boldsymbol{X}\boldsymbol{\beta}]_i$ is SSVS. The above distribution is not, in most cases, of known form but samples can be generated using Gilks and Wild (1992) algorithm as described by Dellaportas and Smith (1993).

2. In most cases ϕ are known. In the case that ϕ is an unknown to be estimated parameter then we generate it from the full conditional posterior distribution

$$f(\phi|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma}, \boldsymbol{y}) \propto exp\left\{ \sum_{i=1}^{n} \frac{y_i g^*(\eta_i) - b\{g^*(\eta_i)\}}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, a_i(\phi)) \right\} f(\phi).$$

3. Generate all pseudo-parameters from the pseudopriors or proposal distributions.

4. Generate the variable indicators γ_j according to one of the following procedures:

   (a) For all $j = 1, \ldots, p$ propose a change from γ_j to $1-\gamma_j$ with probability $q_2(\gamma_j, 1-\gamma_j)$ and accept the proposed move with probability

$$\alpha = min\left(1, \frac{q_2(1-\gamma_j, \gamma_j)}{q_2(\gamma_j, 1-\gamma_j)} O_j^{1-2\gamma_j}\right).$$

   (b) Randomly select a new proposed model $\boldsymbol{\gamma}' \in nb(\boldsymbol{\gamma})$ with probability $j(\boldsymbol{\gamma}, \boldsymbol{\gamma}')$. In most cases this is equivalent to select a term j with probability $q_1(\boldsymbol{\gamma}, j)$ and propose a change from γ_j to $1-\gamma_j$ with probability $q_2(\gamma_j, 1-\gamma_j)$. The acceptance probability is now given by $j(\boldsymbol{\gamma}, \boldsymbol{\gamma}') = q_1(\boldsymbol{\gamma}, j) q_2(\gamma_j, 1-\gamma_j)$ where j is the term in which $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'$ differ. Then accept the proposed move with probability

$$\alpha = min\left(1, \frac{j(\boldsymbol{\gamma}', \boldsymbol{\gamma})}{j(\boldsymbol{\gamma}, \boldsymbol{\gamma}')} O_j^{1-2\gamma_j}\right).$$

The quantity O_j used above is the full conditional posterior odds to include j term in the model given by the ratio

$$O_j = \frac{f(\gamma_j = 1|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}{f(\gamma_j = 0|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma}_{\setminus j}, \boldsymbol{y})}$$

and can be analysed as a product of a likelihood ratio (LR_j), a prior density ratio (PR_j), a pseudoprior (or proposal) density ratio (PSR_j) and the prior model odds $f(\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})$ / $f(\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})$. For more details see Section 4.4 and the Sections describing the corresponding model and variable selection samplers.

4.6.1 Normal Linear Models

Normal linear models are the simplest and most frequently type of generalised linear models. The model formulation is given is given by

$$\boldsymbol{Y} \sim N\left(\boldsymbol{\eta}, \boldsymbol{I}_n\sigma^2\right).$$

In normal models usually we assume common and unknown scale parameter such that $a_i(\phi) = \sigma^2$. Two type of prior setup may be used. The first is the conjugate normal inverse gamma distribution given by (3.6) and (3.7). When we adopt this prior set up then fast variable selection methods described in Section 3.5 should be adopted. The Smith and Kohn (1996) type of prior is proposed. In special cases of orthogonal design matrices (e.g. analysis of variance models with sum to zero constraints) then the samplers proposed by Clyde *et al.* (1996) and or the extension proposed in Section 3.5.1 may be used. In the second case the usual inverse gamma prior on residual variance (or gamma on residual precision; see equation 3.6) can be used but the model parameters do not depend on σ^2. In these cases reversible jump or more complicated Gibbs sampler should be adopted.

Provided that we use the prior distribution (3.3) for model coefficients and independent proposal densities of type (4.12), the sampling procedure for the model parameters in all MCMC variable selection algorithms (except SSVS) are as following

1. Generate $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ from the $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\sigma^2, \gamma, \boldsymbol{y} \sim N\left(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{\gamma})}, \tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})}\right)$ with mean

$$\tilde{\boldsymbol{\beta}}_{(\boldsymbol{\gamma})} = \tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})}\left(\sigma^{-2}\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\boldsymbol{\mu}_{\beta_{(\boldsymbol{\gamma})}}\right)$$

   and covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})} = \left(\sigma^{-2}\boldsymbol{X}^T_{(\boldsymbol{\gamma})}\boldsymbol{X}_{(\boldsymbol{\gamma})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma})}\right)^{-1}.$$

2. Generate $\tau = \sigma^{-2}$ from $G\left(a_\tau + n/2, b_\tau + \left(\boldsymbol{y} - \boldsymbol{X}_{(\boldsymbol{\gamma})}\boldsymbol{\beta}_{(\boldsymbol{\gamma})}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}_{(\boldsymbol{\gamma})}\boldsymbol{\beta}_{(\boldsymbol{\gamma})}\right)/2\right)$.

If SSVS is preferred with prior distribution given in Section 3.4.1.2 then the generation of the model parameters is different since

1. Generate the full parameter vector $\boldsymbol{\beta}$ from the $\boldsymbol{\beta}|\sigma^2, \gamma, \boldsymbol{y} \sim N\left(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{\gamma})}, \tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})}\right)$ with mean

$$\tilde{\boldsymbol{\beta}}_{(\boldsymbol{\gamma})} = \sigma^{-2}\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})}\boldsymbol{X}^T\boldsymbol{y}$$

and covariance matrix

$$\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma})} = \left(\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + (\boldsymbol{D}_{\boldsymbol{\gamma}}\boldsymbol{R}_{\boldsymbol{\gamma}}\boldsymbol{D}_{\boldsymbol{\gamma}})^{-1}\right)^{-1}.$$

2. Generate $\tau = \sigma^{-2}$ from $G\left(a_\tau + n/2, b_\tau + (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})/2\right)$.

4.6.1.1 Simulated Regression Examples

To evaluate the performance of the methods, we use a series of simulated linear regression examples, as presented by Raftery *et al.* (1997). The regression model can be written as $\boldsymbol{y} \sim N(\boldsymbol{\eta}, \boldsymbol{I}\sigma^2)$ with $\boldsymbol{\eta}$ given by (3.2). In all examples independent N(0, 100) priors were used for the regression coefficients and $G(10^{-4}, 10^{-4})$ for the residual precision, σ^{-2}. Finally, we used uniform prior of model space given by $f(\gamma_j = 1) = \pi_j = 0.5$ for all $j \in \mathcal{V}$. The data generation details are given in Table 4.2. For all variable selection procedures we also included the constant term (noted by X_0) as a possible regressor. For SSVS we used the same prior as above for variables included in the model and reduced the variance by a factor of $k = 1000$, for each variable 'excluded' from the model.

			Design	Generated Model		Supported Model	
Dataset	n	p	Structure	$\boldsymbol{\eta}$	σ	Backward/Forward	MCMC
1	50	15	1	$X_4 + X_5$	2.50	$X_4 + X_5 + X_{12}$	$X_4 + X_5$
2	50	15	2	$\sum_{j=1}^5 X_j$	2.50	$\sum_{j=1}^5 X_j + X_{12}/\ X_{14}$	X_{14}
3	50	15	1	0	2.50	*Empty*	*Empty*
4	50	15	2	0	2.50	$X_3 + X_{12}$	X_3
5	100	50	1	0	1.00	X_{19}	*Empty*
6	100	30	1	$0.5X_1$	0.87	X_1	X_1

Table 4.2: Simulated Regression Datasets Details (n is the sample size, p is the number of variables considered excluding the constant term. Design structure 1: $X_0 = 1$, $X_j \sim N(0, 1)$, for $j = 1, \ldots, p$. Design structure 2: $X_0 = 1$, $X_j \sim N(0, 1)$, for $j = 1, \ldots, 10$ and $X_j \sim N(0.3X_1 + 0.5X_2 + 0.7X_3 + 0.9X_4 + 1.1X_5, 1)$, for $j = 11, \ldots, 15$).

The proposal distributions, needed for the implementation of Gibbs variable selection, reversible jump and the Metropolised Carlin and Chib method, were constructed from the sample mean and standard deviation of an initial Gibbs sample run of size 500 of the full model, with initial values taken as zero. To compare the performance of all methods we divided the sample output taken at fixed time intervals (5, 15 and 10 minutes for datasets 1-4, 5 and 6 respectively) into 30 equal batches and reported in Table 4.3 the batch standard deviation of the highest posterior model probability. The evolution of the corresponding ergodic posterior probabilities is displayed in Figures 4.1 and 4.2.

Table 4.4 presents posterior model probabilities estimated by all MCMC methods, as well as, the values of adjusted R^2 and maximum likelihood estimates of residual variance. Valuable information may be also extracted by the marginal posterior variable probabilities presented in Table 4.5. The corresponding p-values when we fit the full model are also presented for comparison purposes. Both Tables 4.4 and 4.5 were constructed after 100,000 iterations for Gibbs variable selection, 1,000,000 iterations for reversible jump and 10,000,000 iterations for Kuo and Mallick, Metropolised Carlin and Chib and SSVS and additional burn-in of ten thousand iterations for all methods.

	Dataset					
	1	2	3	4	5	6
GVS	0.017	0.077	0.024	0.016	0.041	0.027
KM	0.039	0.059	0.032	0.037	0.089	0.059
RJ	0.042	0.102	0.032	0.028	0.062	0.062
MCC	0.044	——	0.043	0.026	0.143	0.078
SSVS	0.138	0.122	0.065	0.111	0.109	0.138

Table 4.3: Simulated Regression Datasets: Batch Standard Deviations of Highest Posterior Model Probability.

As would be expected, all methods gave similar results after a reasonably long run. Generally, Gibbs variable selection seems to have lower batch mean standard deviation than the rest algorithms which indicates greater efficiency. Metropolised Carlin and Chib method seems to have slower rates of convergence and demands more time because in each step (4.7) requires to propose many new values. Note that in the multi-collinear problem (dataset 2), the Metropolised Carlin and Chib sampler did not visit the model selected by the other models. This is not due to the sampler itself but due to the naive proposal pseudoprior setup used that does not exploit the advantages of this sampler; see Section 4.5.1. However, after ten million iterations this method did eventually support the same model. Kuo and Mallick's method generally performs worse than Gibbs variable selection but reasonably well in general, and only in dataset 5 it has considerably higher standard deviation than all the other methods. Finally, SSVS has systematically higher standard deviations than all the other methods.

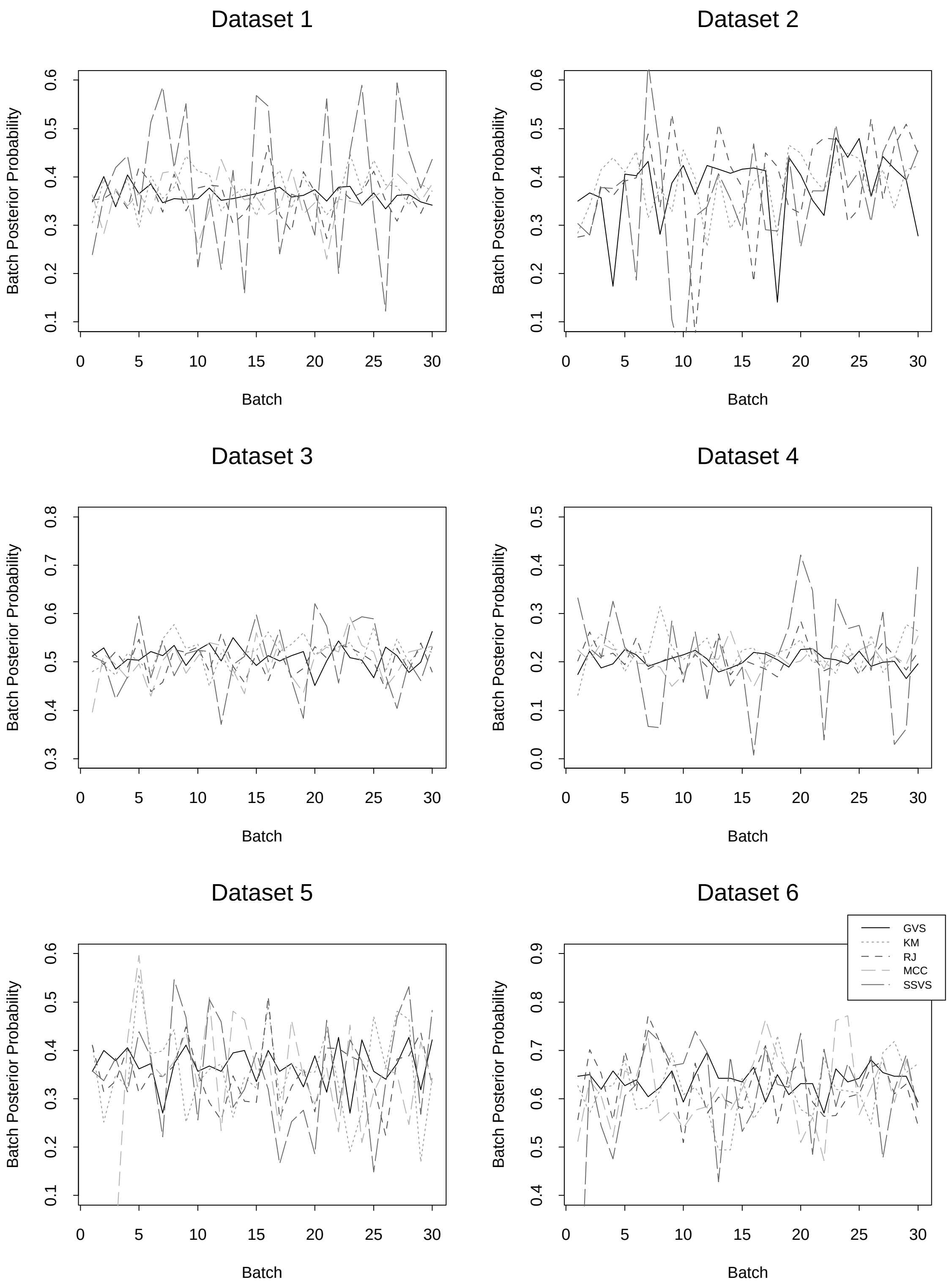


Figure 4.1: Simulated Regression Datasets: Batch Highest Posterior Model Probabilities (GVS: Gibbs variable selection, KM: Kuo and Mallick sampler, RJ: reversible jump, MCC: Metropolised Carlin and Chib SSVS: stochastic search variable selection).

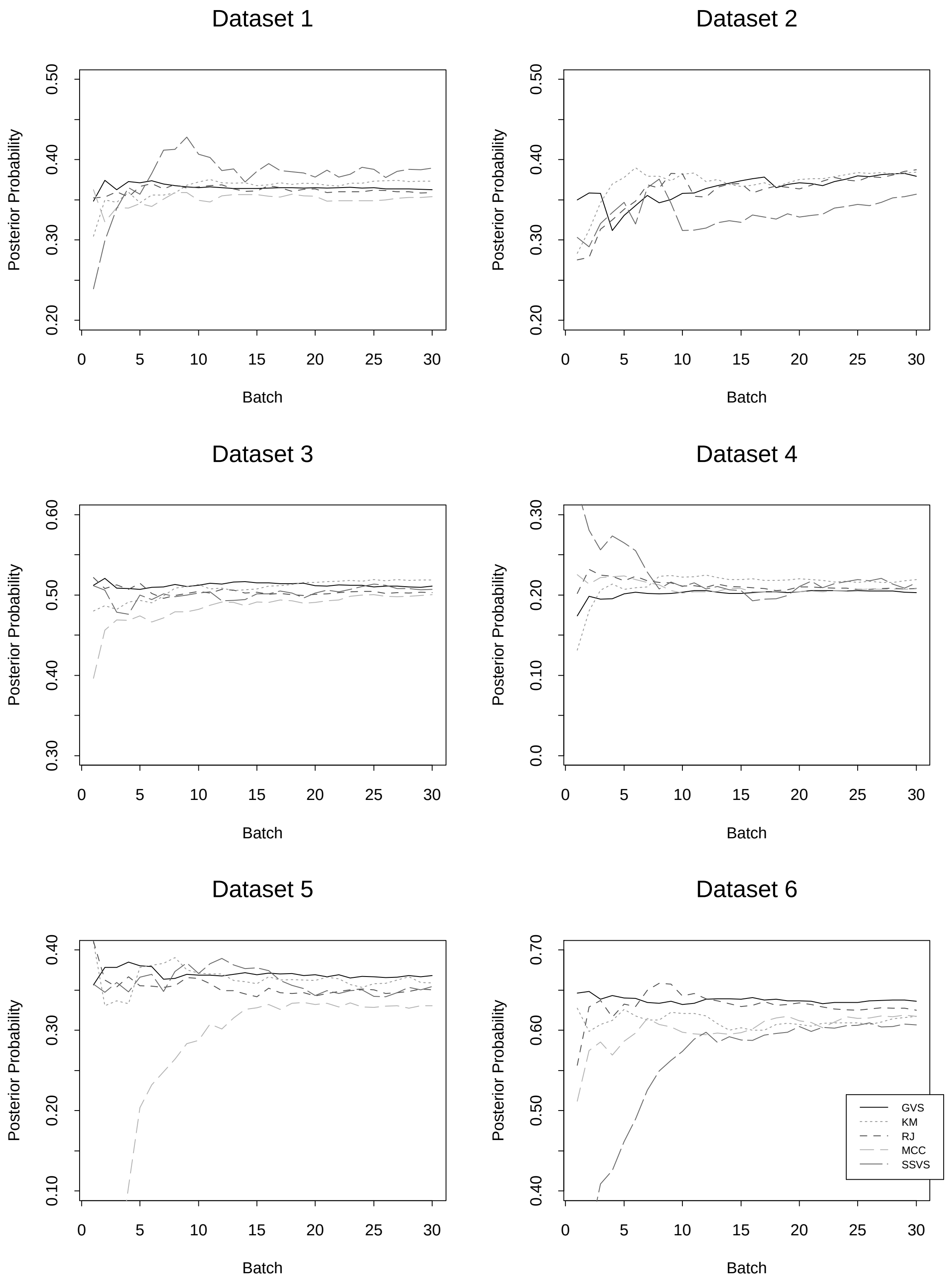


Figure 4.2: Simulated Regression Datasets: Ergodic Highest Posterior Model Probabilities.

		Posterior Probabilities of MCMC Algorithms					ML Measures	
Dataset	Model	GVS	KM	RJ	MCC	SSVS	R^2_{adj}	$\hat{\sigma}$
1	$X_4 + X_5$	0.358	0.371	0.358	0.363	0.361	0.468	2.44
	$X_4 + X_5 + X_{12}$ **	0.185	0.180	0.186	0.183	0.180	0.507	2.35
	Models with $p < 0.05$	0.457	0.449	0.456	0.454	0.459		
2	X_{14} *	0.383	0.375	0.380	0.388	0.380	0.452	2.73
	$X_{13} + X_{14}$	0.052	0.052	0.056	0.055	0.053	0.476	2.67
	Models with $p < 0.05$	0.565	0.573	0.564	0.557	0.567		
3	*Empty* **	0.504	0.509	0.504	0.503	0.500	—	2.24
	X_1	0.057	0.058	0.060	0.059	0.056	0.033	2.20
	Models with $p < 0.05$	0.439	0.433	0.436	0.438	0.444		
4	X_3	0.207	0.213	0.209	0.211	0.204	0.165	2.65
	$X_3 + X_{12}$ **	0.098	0.099	0.097	0.096	0.096	0.221	2.56
	Models with $p < 0.05$	0.695	0.688	0.694	0.693	0.700		
5	*Empty*	0.373	0.374	0.366	0.373	0.375	—	0.98
	X_{19} **	0.073	0.071	0.072	0.072	0.071	0.045	0.96
	X_{38}	0.057	0.059	0.058	0.056	0.059	0.071	0.94
	Models with $p < 0.05$	0.497	0.496	0.504	0.499	0.495		
6	X_1 **	0.621	0.623	0.613	0.618	0.619	0.235	0.79
	$X_1 + X_{11}$	0.064	0.064	0.065	0.064	0.062	0.266	0.77
	Models with $p < 0.05$	0.315	0.313	0.322	0.318	0.319		

Table 4.4: Simulated Regression Datasets: Posterior Model Probabilities (* = selected by forward procedure, ** = selected by both forward and backward procedures).

		Posterior Probabilities of MCMC Algorithms					Full Model
Dataset	Model	GVS	KM	RJ	MCC	SSVS	P-value
1	X_5	1.000	1.000	1.000	1.000	1.000	0.000
	X_4	0.969	0.967	0.970	0.969	0.969	0.001
	X_{12}	0.341	0.327	0.340	0.337	0.339	0.041
	X_{15}	0.078	0.073	0.075	0.078	0.077	0.102
	X_{11}	0.050	0.048	0.051	0.049	0.051	0.338
2	X_{14}	0.902	0.903	0.907	0.944	0.915	0.788
	X_{13}	0.158	0.158	0.160	0.157	0.153	0.591
	X_1	0.127	0.122	0.127	0.124	0.123	0.002
	X_{11}	0.084	0.087	0.084	0.078	0.083	0.338
	X_{12}	0.072	0.075	0.064	0.057	0.072	0.041
	X_3	0.064	0.061	0.066	0.065	0.061	0.002
	X_2	0.053	0.052	0.054	0.040	0.052	0.010
	X_{10}	0.051	0.051	0.051	0.050	0.052	0.439
3	X_1	0.098	0.100	0.100	0.101	0.098	0.934
	X_2	0.080	0.075	0.079	0.080	0.082	0.070
	X_9	0.060	0.060	0.061	0.059	0.060	0.204
	X_{10}	0.057	0.055	0.056	0.059	0.059	0.653
	X_6	0.050	0.051	0.050	0.051	0.051	0.806
4	X_3	0.810	0.810	0.812	0.812	0.804	0.031
	X_{12}	0.322	0.316	0.324	0.319	0.323	0.094
	X_1	0.154	0.145	0.154	0.153	0.157	0.226
	X_0	0.123	0.118	0.121	0.122	0.122	0.204
5	X_{19}	0.147	0.146	0.152	0.146	0.145	0.036
	X_{38}	0.127	0.127	0.131	0.125	0.126	0.195
	X_{46}	0.073	0.072	0.072	0.072	0.069	0.623
6	X_1	1.000	1.000	1.000	1.000	1.000	0.000
	X_{11}	0.091	0.091	0.094	0.091	0.089	0.126

Table 4.5: Simulated Regression Datasets: Posterior Variable Probabilities Higher than 0.05.

4.6.2 Poisson Models

This section provides details of how to implement model selection methods in Poisson log-linear models. Results of this section are also given in two research papers; see Ntzoufras *et al.* (1996) for a comparison of MCMC model selection algorithms in log-linear models and Ntzoufras *et al.* (1998) for the implementation of SSVS in log-linear models.

The model formulation is given is

$$Y_i \sim Poisson\left(e^{\eta_i}\right).$$

The linear predictor η_i is generally defined by equation (4.15) and by (4.14) for Gibbs variable selection and (4.16) for SSVS. Therefore the likelihood is given by

$$f(\boldsymbol{y}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})},\gamma) = exp\left\{-\sum_{i=1}^{n} e^{\eta_i} + \sum_{i=1}^{n} y_i\eta_i - \sum_{i=1}^{n} log(y_i!)\right\}.$$

Provided that we use the prior distribution (3.3) for model coefficients and independent proposal densities of type (4.12), the sampling procedure for the model parameters for the MCMC variable selection algorithms are as following

1. Generate $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ from equation (4.17) in Carlin and Chib variants, (4.18) in Gibbs variable selection and (4.19) in SSVS. The first equation can also be used in Gibbs variable selection without any complication. $f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\gamma,\boldsymbol{y}) \propto$

$$(RJ/CCVS) \quad \propto \quad exp\left\{-\sum_{i=1}^{n} e^{\left[\boldsymbol{X}_{(\gamma)}\boldsymbol{\beta}_{(\gamma)}\right]_i} + \sum_{i=1}^{n} y_i\left[\boldsymbol{X}_{(\gamma)}\boldsymbol{\beta}_{(\gamma)}\right]_i\right\} f\left(\boldsymbol{\beta}_{(\gamma)}|\boldsymbol{\gamma}\right), \qquad (4.17)$$

$$(GVS) \quad \propto \quad exp\left\{-\sum_{i=1}^{n} exp\left(\sum_{j\in\mathcal{V}}\gamma_j\boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right) + \sum_{i=1}^{n}\sum_{j\in\mathcal{V}} y_i\gamma_j\boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right\} f\left(\boldsymbol{\beta}_{(\gamma)}|\boldsymbol{\gamma}\right), \qquad (4.18)$$

$$(SSVS) \quad \propto \quad exp\left\{-\sum_{i=1}^{n} exp\left(\sum_{j\in\mathcal{V}}\boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right) + \sum_{i=1}^{n}\sum_{j\in\mathcal{V}} y_i\boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right\} f\left(\boldsymbol{\beta}|\boldsymbol{\gamma}\right), \qquad (4.19)$$

   where $\boldsymbol{X}_{ij}$ is the $1\times d_j$ sub-matrix corresponding to i observation and j term.

2. The scale parameter is known and therefore we do not have to estimate it.

4.6.2.1 SSVS Prior Distributions for Contingency Tables Problems with Two Leveled Factors

Here we introduce an approach similar to the one described in Section 3.4.1.2 to define sensible prior distributions for implementing SSVS in Poisson log-linear models. For log-linear models the specification of Σ_j and k_j is more straightforward than for linear regression

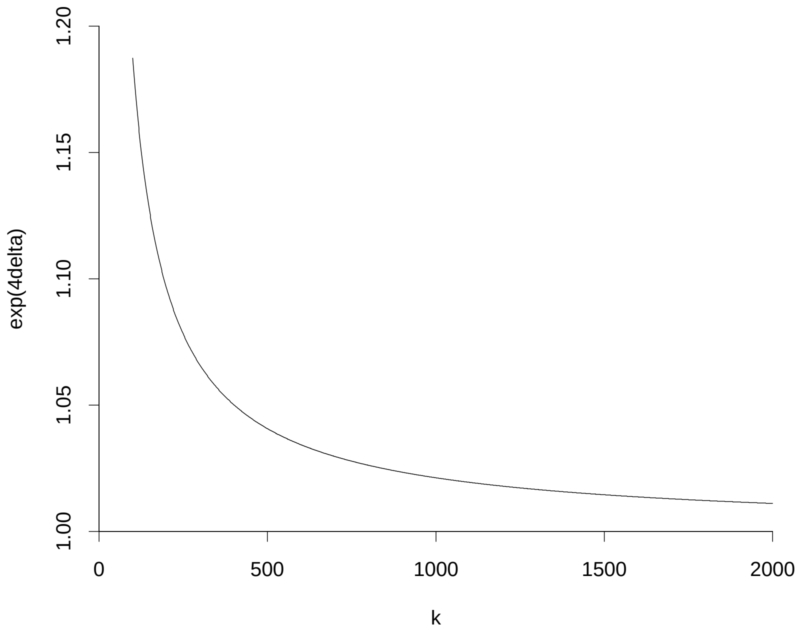


Figure 4.3: The Relationship between Cross-product Ratio Boundary ($= e^{4\delta}$) and k for the 2×2 Table

models, as there is no intrinsic unknown scale parameter σ^2 and hence parameter values have the same interpretation in different examples. We adopt the prior of Dellaportas and Forster (1999) for which $Var(\boldsymbol{\beta}_j) = \boldsymbol{\Sigma}_j = c_j^2 \boldsymbol{V}_j$, where $\boldsymbol{V}_j$ is given by (3.5). When $\boldsymbol{\beta}_j$ is one-dimensional it is straightforward to define the parameters k_j. Consider the simplest possible example, the 2×2 contingency table. When j is the interaction between the two binary factors, then $d_j = 1$, $V_j = (1/d)$ (or $\Sigma_j = c_j^2/d$) and β_j is equal to one quarter of the log cross-product ratio. Now, suppose that δ_j is the smallest value of β_j of practical significance. Then

$$\delta_j \quad = \quad \sqrt{c_j^2 \frac{2\log k_j}{d(k_j^2 - 1)}} \quad \approx \quad c_j k_j^{-1} \sqrt{\frac{2 \log k_j}{d}}.$$

Therefore the prior may be constructed by specifying δ_j and either k_j or c_j. If, as suggested above, $c_j^2 = 2d$, then the prior is specified through

$$\delta_j \quad = \quad 2\sqrt{\frac{\log k_j}{k_j^2 - 1}} \quad \approx \quad 2k_j^{-1}\sqrt{\log k_j}.$$

When the model consists completely of one-dimensional parameters, this approach can be adopted for every model term j since each parameter is associated with a corresponding odds ratio. Indeed, for a reference analysis, it may be appropriate to choose the same values for k_j for every term j. For example, if $k_j = 10^3$ and $c_j^2 = 2d = 8$, then $\Sigma_j = 2$ as suggested by Dellaportas and Forster (1999), and $e^{4\delta_j} \approx 1.021$. So, the boundary between significant and insignificant cross product ratios would be represented by an increase in 'risk' of around 2.1%. The relationship between k and δ for the 2×2 reference example is illustrated in Figure 4.3.

4.6.2.2 A Large 2^6 Contingency Table Example

Consider the 2^6 table of risk factors for coronary heart disease presented by Edwards and Havránek (1985). This table has also been analysed by Madigan and Raftery (1994) and Madigan *et al.* (1995) using both stepwise and MCMC Bayesian model selection algorithms for decomposable log-linear models. Decomposable models are a subset of the hierarchical models which we consider in this example. However, there are many interesting models which are not decomposable, and therefore we present the results of our SSVS hierarchical model selection approach. Here, the six variables are: A, smoking; B, strenuous mental work; C, strenuous physical work; D, systolic blood pressure; E, ratio of α and β lipoproteins; F, family anamnesis of coronary heart disease; see Table 4.6.

			C	No				Yes			
			B	No		Yes		No		Yes	
F	E	D	A	No	Yes	No	Yes	No	Yes	No	Yes
Negative	< 3	< 140		44	40	112	67	129	145	12	23
		≥ 140		35	12	80	33	109	67	7	9
	≥ 3	< 140		23	32	70	66	50	80	7	13
		≥ 140		24	25	73	57	51	63	7	16
Positive	< 3	< 140		5	7	21	9	9	17	1	4
		≥ 140		4	3	11	8	14	17	5	2
	≥ 3	< 140		7	3	14	14	9	16	2	3
		≥ 140		4	0	13	11	5	14	4	4

Table 4.6: 2^6 Contingency Table: Edwards and Havránek (1985) Dataset.

The posterior distribution of $\boldsymbol{\gamma}$ is summarised in Table 4.7 by presenting the models corresponding to the most probable $\boldsymbol{\gamma}$ together with the corresponding posterior probability $f(\boldsymbol{\gamma}|\boldsymbol{y})$.

Model	Posterior Probability			
	RJ	KM	GVS	SSVS
AC+BC+AD+AE+CE+DE+F	0.260	0.268	0.270	0.270
AC+BC+AD+AE+BE+DE+F	0.166	0.157	0.154	0.164
AC+BC+AD+AE+CE+DE+BF	0.075	0.073	0.077	0.070
AC+BC+AD+AE+BE+CE+DE+F	0.063	0.073	0.069	0.054
Other Models	0.460	0.429	0.430	0.443
Mean of Batch Standard Deviation	0.064	0.039	0.025	0.107

Table 4.7: 2^6 Contingency Table: Posterior Model Probabilities.

We use the same $N(0, \Sigma)$ prior distribution for each $\beta_j,\ j = 1, \ldots, 64$, under all possible models. We use the prior of Dellaportas and Forster (1999) resulting in $N(0, 2)$ for each all $\beta_j,\ j = 1, \ldots, 64$ In SSVS we used the same 'large' variance while the 'small' one was reduced by $k_j = 1000$ for all model parameters in order to give approximately the same results as the other methods. As model space $\mathcal{M}$ we consider the set of hierarchical models for contingency tables. Also note that the terms involving single factors were always included in the model since we are interested in the association between the six factors used. A uniform prior on the model space of hierarchical log-linear models was adopted.

All Markov chains were initiated at the full model with starting points $\beta_j = 0$ for all $j = 1, \ldots, 64$. For the reversible jump methods we always propose a 'neighbouring' model which differs from the current model by one term, hence $j(m, m) = 0$. Within each model, updating of the parameters β_j was performed via Gibbs sampling steps as described in Dellaportas and Smith (1993). Finally, each Markov chain ran for 110,000 iterations and the output summaries are based on ergodic averages taken on the last 100,000 iterations.

All samplers were extremely mobile, and as would be hoped, gave similar results of the model probabilities in a reasonably short time. SSVS with $k = 1000$ gave similar results to the other model selection methods while the choice of $k = 100$ resulted in support of different models (see Figure 4.7). The full results are given in Table 4.7. The proposal distributions for β_j in reversible jump as well as the pseudopriors in Gibbs variable selection of Table 4.7 are $N(0, \Sigma/100)$. This proposal worked better than using pseudopriors from a pilot run of the full model mainly due to the small standard error of these estimates. An alternative to use variances of type Σ/k^2 for various values of k. It is interesting to note that the two models with highest probability are the same as those determined by Edwards and Havránek (1985) using their procedure for identifying acceptable parsimonious models. Furthermore, none of these models are decomposable, and hence they were not identified by Madigan and Raftery (1994) or Madigan *et al.* (1995).

Mean values of batch standard deviations are presented as a measure of convergence. Gibbs variable selection has the lowest value while SSVS the largest. Kuo and Mallick sampler performs well enough. Metropolised Carlin and Chib was not used since they are not efficient in such examples.

The resulting 110,000 iterations (discarding the first 10,000) were divided in batches of length of 2,000 iterations. Probabilities of the best four models over different batches are given in barplots (see Figure 4.5) while their ergodic probabilities are given in Figure 4.6. Assessment of convergence can be done via these figures. Smooth changes of models in barplots indicate convergence of the posterior distribution of the model indicator m.

Different values of the proposal parameter k in Gibbs variable selection were used in order to assess which choice leads to faster convergence. From the plots it is obvious that $k = 10$ for automatic proposal works better than the other choices. Densities with maximum likelihood values perform worst than automatic proposals and therefore a calibration of their variance is suggested. Figure 4.4 presents the mean of batch standard deviations in automatic proposal densities for different values of k^2. Optimal values of k (for this example) seem to be close to $k = 10$.

Dellaportas and Forster (1999) type of prior densities were also used with larger $c^2 = c'd$ for $c' = 2, 5, 10, 100$ to assess robustness. The value of $c^2 = n$ ($c' = 28.76$) was also used. In all prior choices used the two supported models using Dellaportas and Forster (1999) originally proposed prior ($c' = 2$) have still the highest posterior probabilities (see Figure 4.8). The fitted lines are produced using a logistic regression model with covariates the

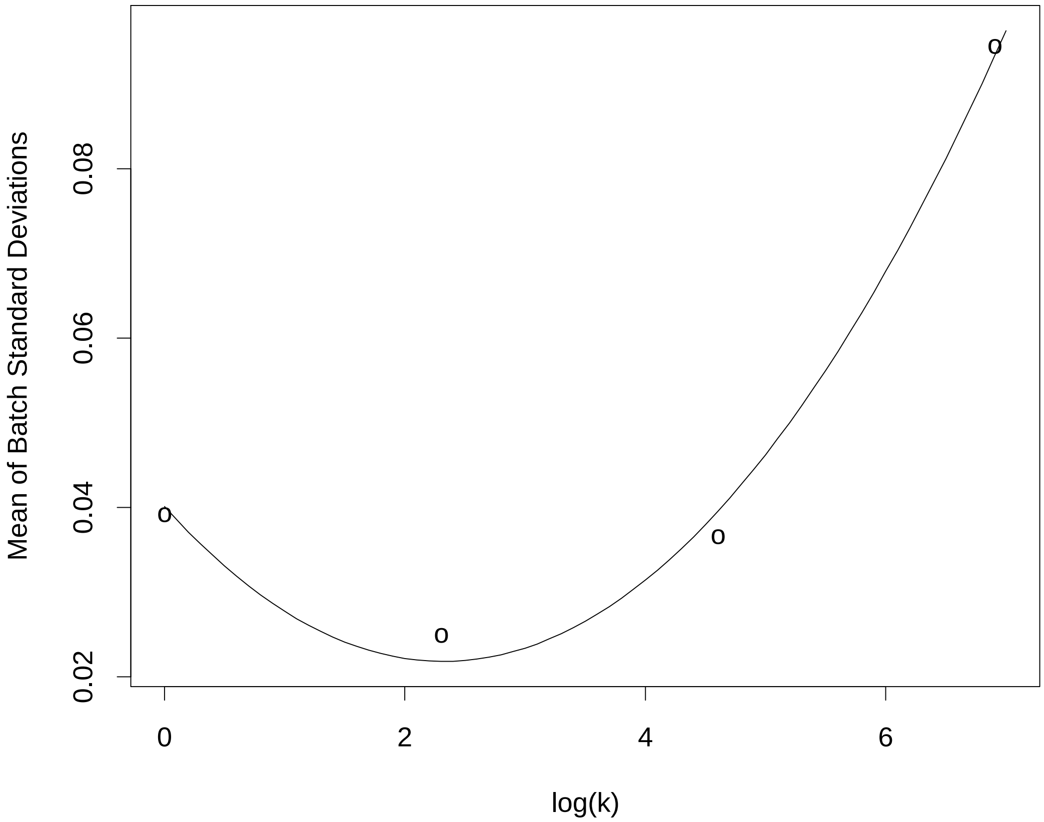


Figure 4.4: 2^6 Contingency Table: Mean Values of GVS Batch Standard Deviations for Different k.

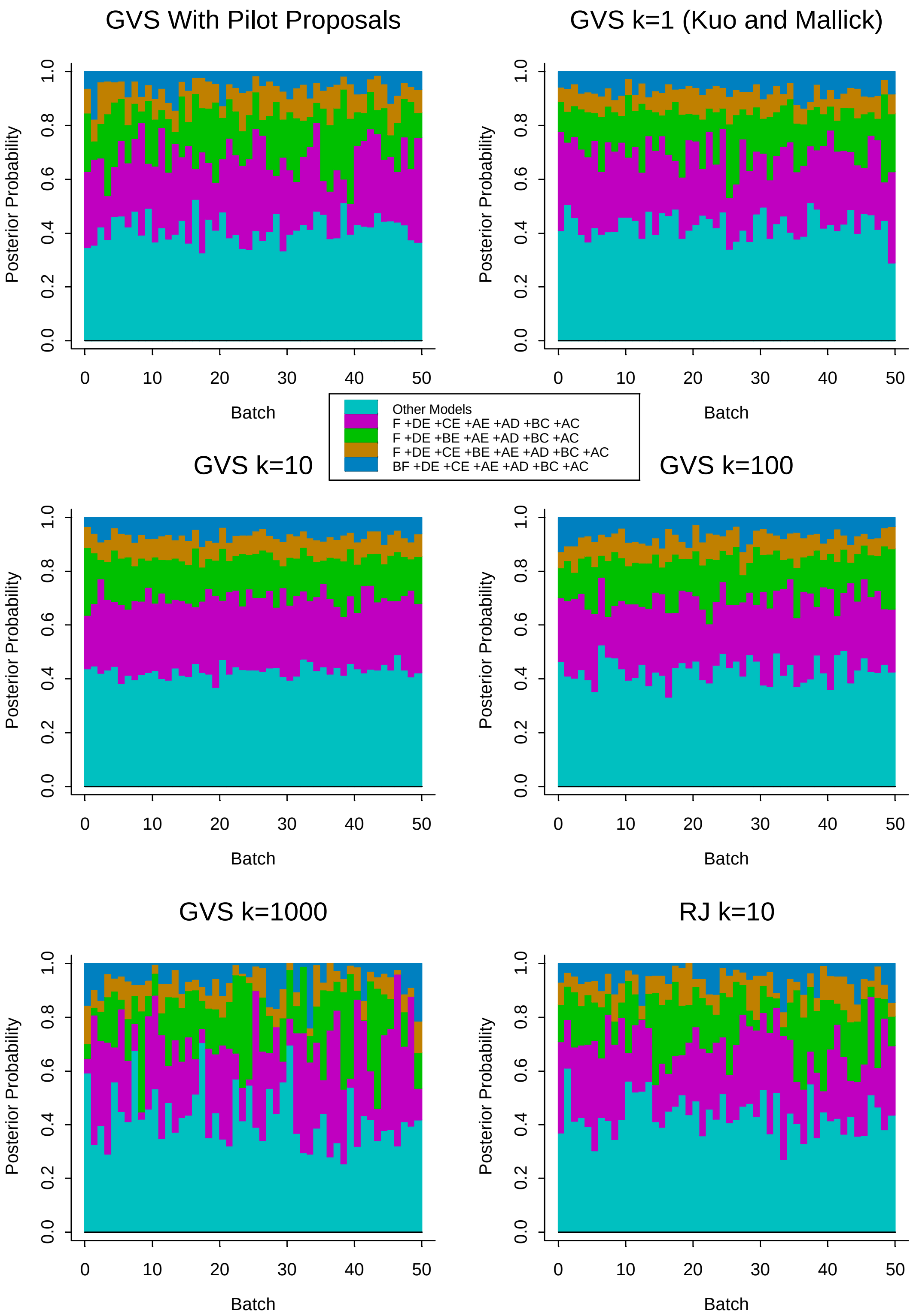


Figure 4.5: 2^6 Contingency Table: Barplot Comparison of Different GVS Proposal Setups.

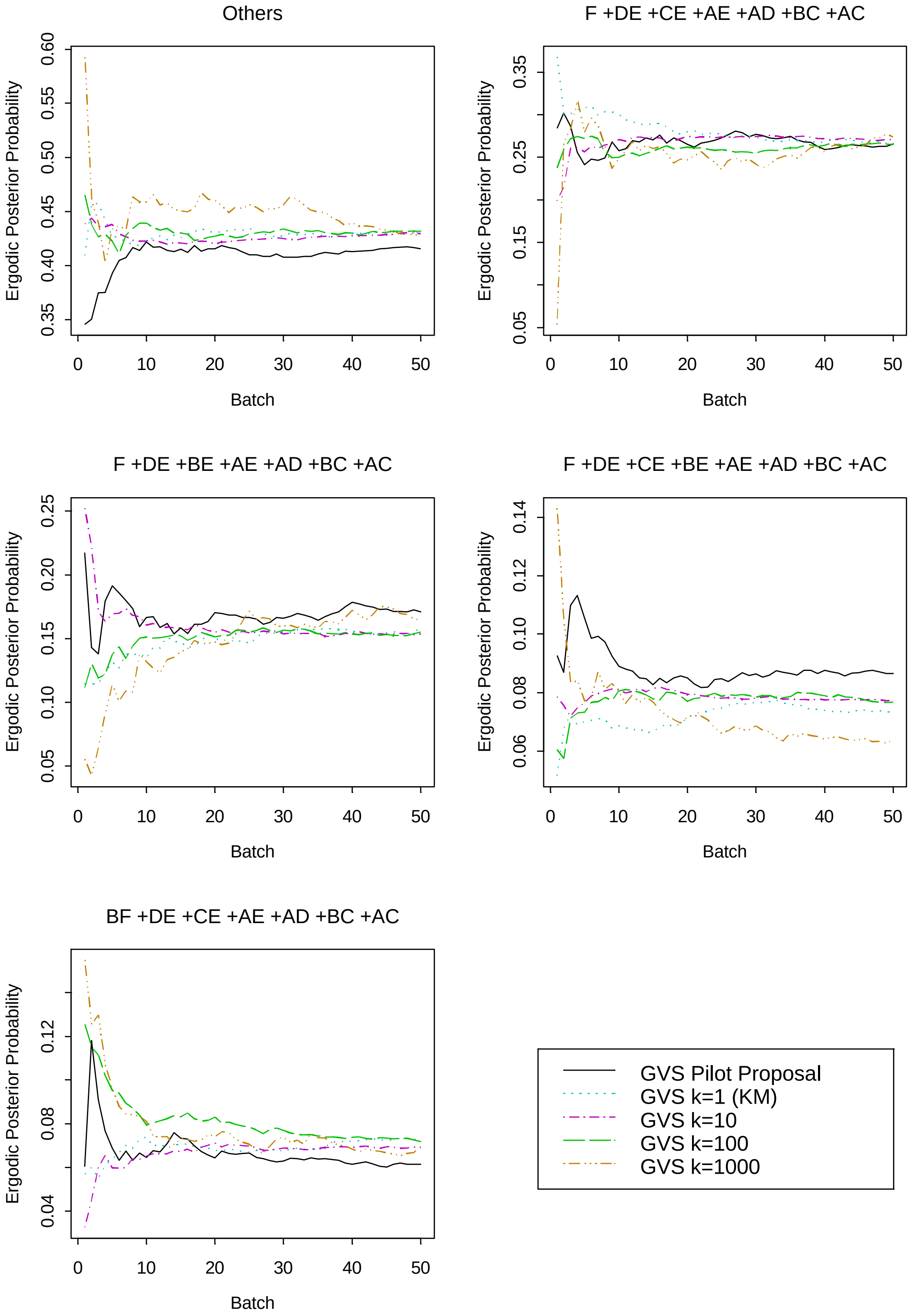


Figure 4.6: 2^6 Contingency Table: Ergodic Posterior Probabilities Comparison of Different GVS Proposal Setups.

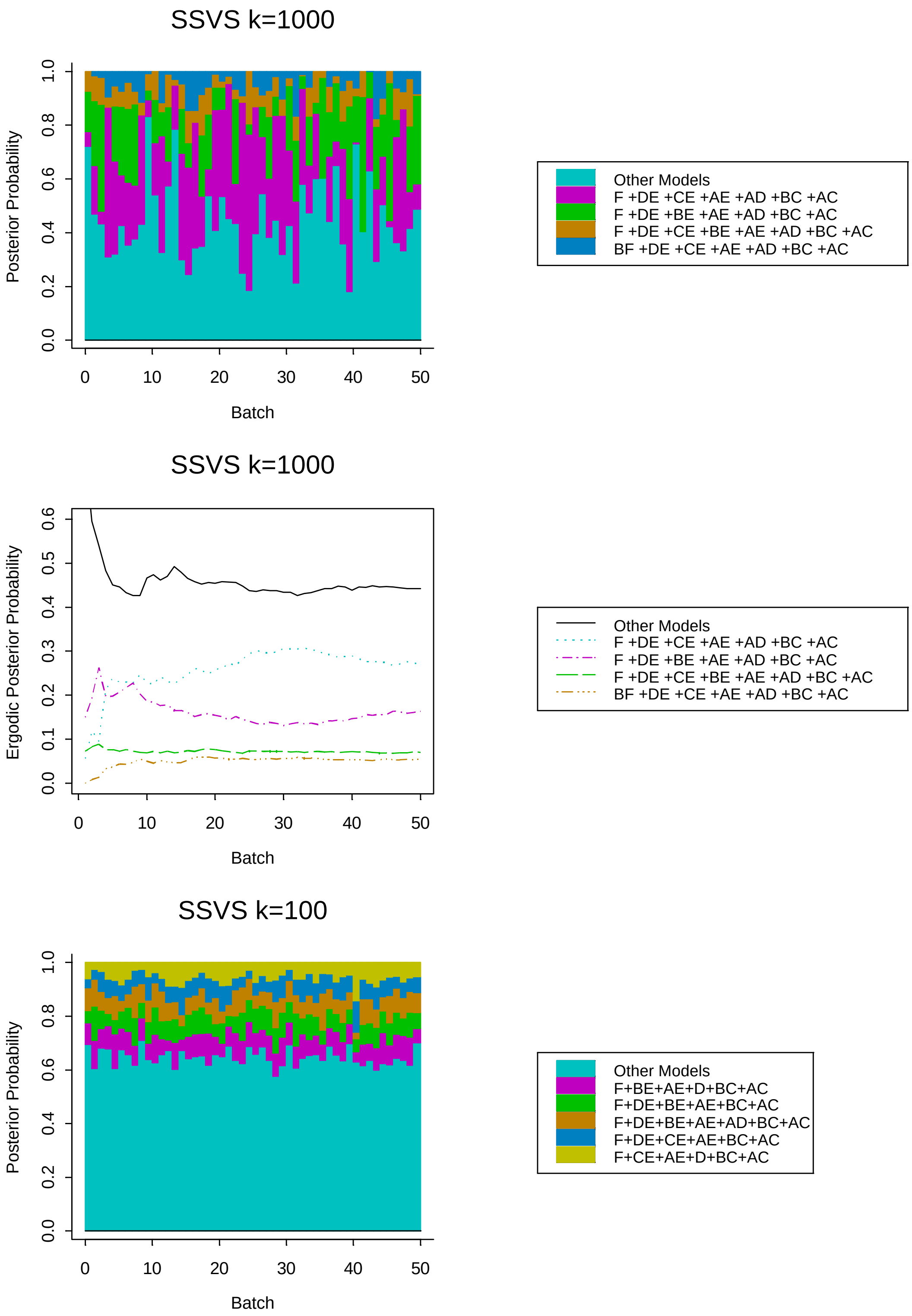


Figure 4.7: 2^6 Contingency Table: Plots for SSVS.

values of c' and its logarithm. This model was used due to the form of the Bayes factor in the normal model when conjugate normal inverse gamma prior was used. The fitted logit model for the posterior probability for model $AC + BC + AD + AE + CE + DE + F$ is given by $log[p/(1-p)] = -1.133 + 0.307 log(c') - 0.012c'$ with maximum achieved at $c' = 24.90$. Similarly the maximum posterior probability for the second best model is achieved for prior with $c' = 24.98$. The posterior probabilities increase until these values and then drop since flatter priors support simplest models. If we extend the graph in larger values the four model probabilities examined here will degenerate to zero. In this example it is very important that, for this range of priors, the models with the two highest posterior probabilities remain the same and for this reason we have an additional strong argument in favour of these two models.

4.6.2.3 SSVS Prior Distribution for Factors with Multiple Categories

One problem that we may face up when using SSVS is what prior distributions we shall adopt when multidimensional terms are involved in the model selection procedure (for example in analysis of variance or contingency table models). In such cases some adjustment is required otherwise the algorithm either will stuck in the more complicated model or will move very slowly.

A first simple approach is to ensure that the ratio of the two components of the mixture prior density at $\boldsymbol{\beta}_j = \mathbf{0}$ is invariant to the dimension d_j of $\boldsymbol{\beta}_j$ by setting $\log k_j$ proportional to $1/d_j$. Then $f(\gamma_j|\boldsymbol{\beta}_j = \mathbf{0})$ will not depend on the dimension d_j of $\boldsymbol{\beta}_j$. This seems to ensure sensible results in problems where the d_j vary, and is intuitively plausible as the interpretation of $\boldsymbol{\beta}_j = \mathbf{0}$ is invariant to the dimension d_j.

An alternative approach for multidimensional $\boldsymbol{\beta}_j$ is to adopt an approach similar to semi-automatic method of George and McCulloch (1993). If we consider the SSVS prior (3.23) with $\boldsymbol{\Sigma}_j = c_j^2 \boldsymbol{V}_j$ the we can choose k_j by considering the values of $\boldsymbol{\beta}_j$

$$\boldsymbol{\beta}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{\beta}_j = 2 d_j c_j^2 \frac{\log k_j}{k_j^2 - 1}$$

where the two components of the mixture prior densities have equal values. Suppose that

$$\mathcal{G}_j = \left\{ \boldsymbol{\beta}_j : \boldsymbol{\beta}_j^T \boldsymbol{V}_j^{-1} \boldsymbol{\beta}_j \leq 2 d_j c_j^2 \frac{\log k_j}{k_j^2 - 1} \right\}$$

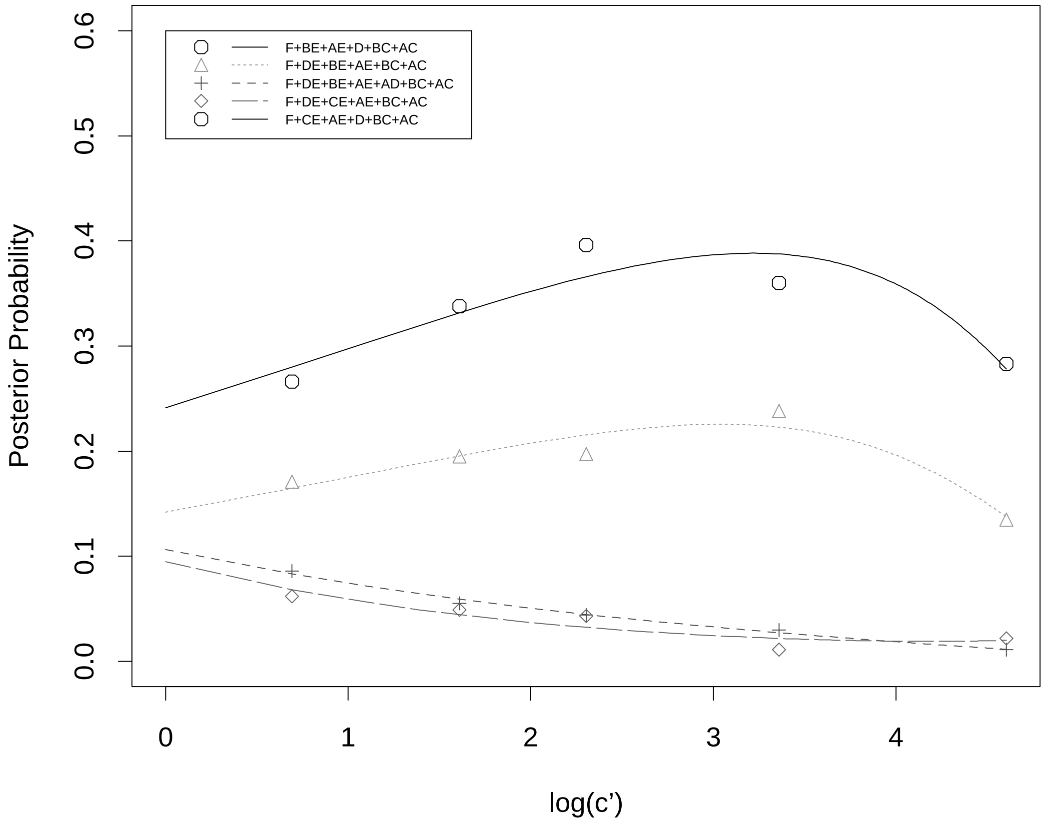


Figure 4.8: 2^6 Contingency Table: Variation of Posterior Model Probabilities for Different Prior Variance.

denotes the 'region of insignificance'. Then it is possible to determine c_j and k_j so that $P(\boldsymbol{\beta}_j \in \mathcal{G}_j|\gamma_j = 0)$ is the same for all j, regardless of the value of d_j. We have

$$P(\boldsymbol{\beta}_j \in \mathcal{G}_j|\gamma_j = 0) \quad = \quad F_{\chi^2_{d_j}}\left(2d_j\frac{k_j^2}{k_j^2-1}logk_j\right)$$

where $F_{\chi^2_d}$ is the distribution function of a chi-squared random variable with d degrees of freedom. Therefore, if k is the value of k_j when $d_j = 1$, then the corresponding value for $d_j > 1$ is given by the solution of the equation

$$F_{\chi^2_{d_j}}\left(2d_j\frac{k_j^2}{k_j^2-1}\log k_j\right) \quad = \quad F_{\chi^2_{d_1}}\left(2\frac{k^2}{k^2-1}\log k\right) \tag{4.20}$$

and approximately,

$$k_j \quad = \quad \exp\left(\frac{1}{2d_j}F^{-1}_{\chi^2_{d_j}}\left[F_{\chi^2_{d_1}}\left(2\log k\right)\right]\right). \tag{4.21}$$

Equivalent arguments can be used for defining priors for multivariate terms in logistic regression and ANOVA models.

To see how k_j varies with dimension, see Figure 4.9, which is a plot of $\log k_j$ against d_j when $k = 1000$. The solid line represents $\log k_j \propto 1/d_j$ while the dotted line represents the values given by (4.21).

4.6.2.4 An Example with Multiple Categories Factors: $3 \times 2 \times 4$ Contingency Table

This example is a $3 \times 2 \times 4$ contingency table presented by Knuiman and Speed (1988) where 491 individuals are classified by three categorical variables: obesity (O: low,average,high), hypertension (H: yes,no) and alcohol consumption (A: 1,1–2,3–5,6+ drinks per day); see Table 4.8.

The results are summarised in Tables 4.9 and 4.10. There are nine possible hierarchical models in total, but the data strongly favour the model of mutual independence of H, O and A, with some evidence of an interaction between O and H.

We use the prior proposed by Dellaportas and Forster (1999). We adopt here the proposal and pseudoprior densities of type $N(0, \boldsymbol{\Sigma}_j/k^2)$ for various choices of k. Pilot run estimates were avoided in order to simplify computations. In SSVS the two adjustment approaches were used for two different 'small' variances ($k = 1000$ and $k = 5000$).

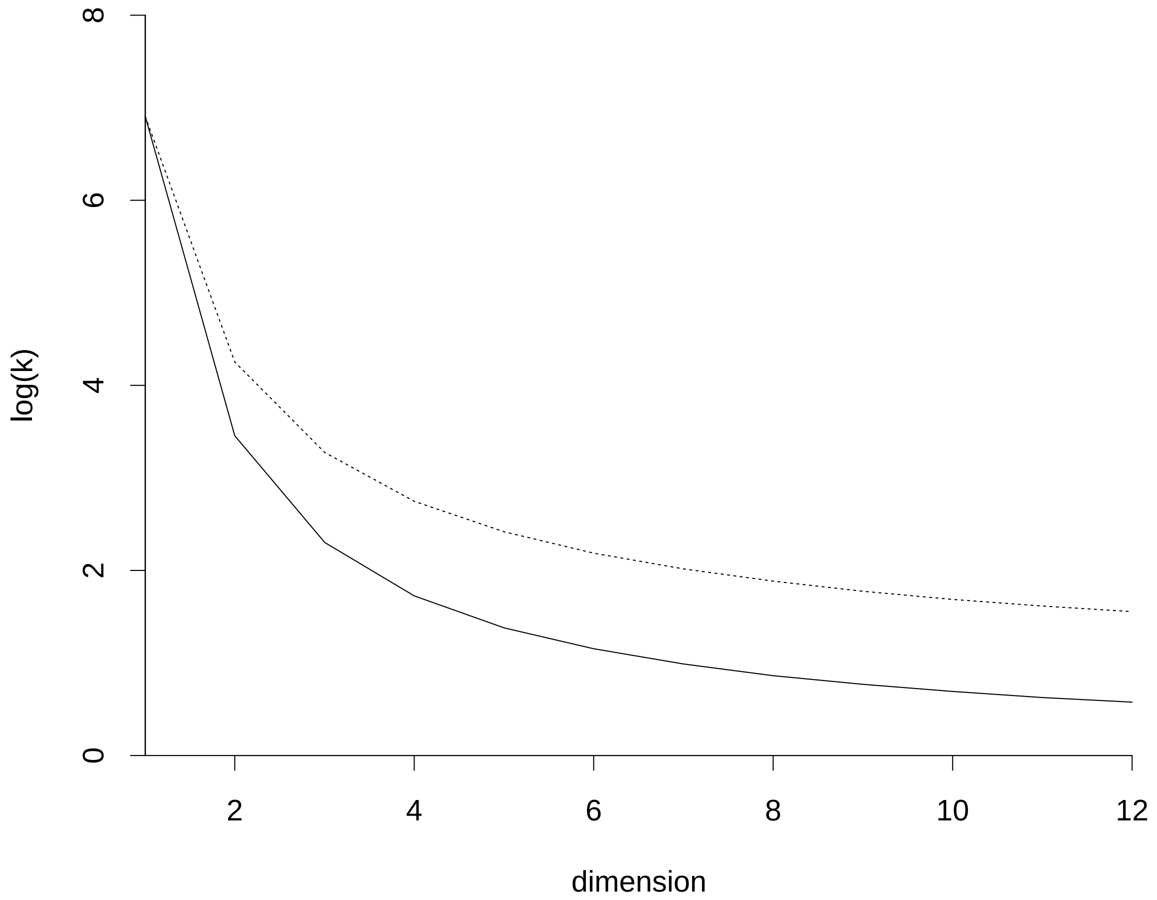


Figure 4.9: The Relationship between $\log k_j$ and d_j for $k = 1000$ (Solid line for $\log k_j \propto 1/d_j$ and Dotted line for equation 4.21).

		Alcohol Intake			
Obesity	**High BP**	0	1-2	3-5	6+
Low	Yes	5	9	8	10
	No	40	36	33	24
Average	Yes	6	9	11	14
	No	33	23	35	30
High	Yes	9	12	19	19
	No	24	25	28	29

Table 4.8: $3 \times 2 \times 4$ Contingency Table: Knuiman and Speed (1988) Dataset.

	Posterior Probability			
	RJ	KM	GVS	GVS
Model	$(k = 10)$	$(k = 1)$	$(k = 10)$	$(k = 100)$
$O + H + A$	0.685	0.688	0.674	0.680
$OH + A$	0.309	0.307	0.320	0.315
Other Models	0.006	0.005	0.006	0.005
Mean of Batch Standard Deviation	0.046	0.129	0.030	0.174

Table 4.9: $3{\times}2{\times}4$ Contingency Table: Posterior Model Probabilities Estimated by Reversible Jump and Gibbs Variable Selection Methods (Proposals and Pseudopriors: $N(\mathbf{0}, \Sigma_j/k^2)$.

	Multivariate Adjustment Criterion			
	[1] $\log c_j \propto 1/\delta_j$		[2] Equation (4.21)	
Model	$k = 1000$	$k = 5000$	$k = 1000$	$k = 5000$
$O + H + A$	0.948	0.853	0.829	0.764
$OH + A$	0.049	0.145	0.169	0.233
Others	0.003	0.002	0.002	0.003
Mean of Batch Standard Deviation	0.012	0.053	0.064	0.233

Table 4.10: $3 \times 2 \times 4$ Contingency Table Example: Posterior Model Probabilities estimated by SSVS.

All methods support the model independence with high probability. Gibbs variable selection methods and reversible jump support model $H + O + A$ with probability about 68%. All SSVS samples support the same model with higher probabilities. Generally the first adjustment method supports more complicated models.

In terms of convergence both reversible jump and Gibbs variable selection with $k = 10$ seem to reach convergence very fast (see Figure 4.11) while Kuo and Mallick and Gibbs variable selection with $k = 100$ demonstrate more variability among the batches. Similar arguments hold for SSVS methods. The chain resulted from SSVS that reached convergence faster in terms of batch standard deviation is the one with $k = 1000$ and the first criterion. This result was expected since the first criterion utilizes smaller values for k_j of multidimensional terms.

Finally, flatter prior distributions were also used to assess the effect on the posterior distribution. We use the variance multiplicator $c^2 = c'd$ ($c' = 2$ results to Dellaportas and Forster prior) for values of $c' \in \{2, 5, 10, 30, 50, 100\}$. Results are presented in Figure 4.10. The fitted line is produced by fitting a logistic regression model on the probability of the best model with regressors c' and $log(c')$. As expected the probability of model $H + O + A$ tends to zero as the variance becomes larger. The supported model even for $c' = 2$ is the simplest possible in the set of models that we consider and for this reason when the prior becomes flatter the probability of the simplest model tends to one.

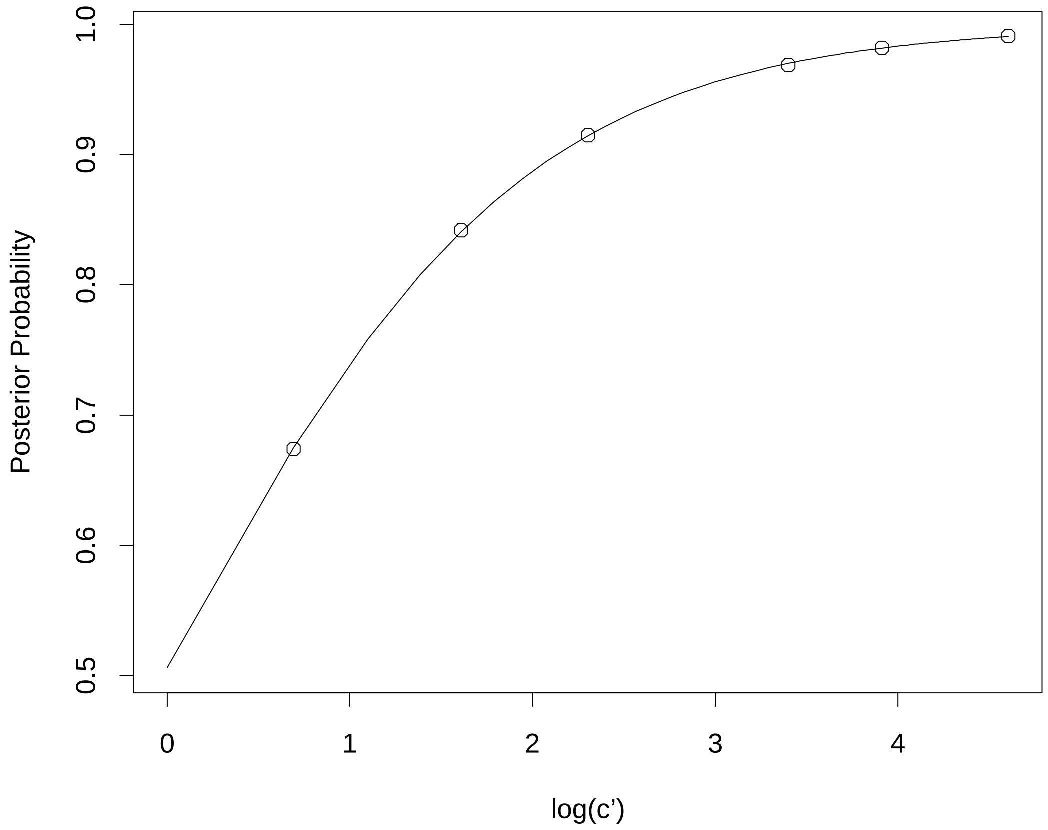


Figure 4.10: $3 \times 2 \times 4$ Contingency Table Example: Variation of Posterior Probability of Model $H + O + A$ for Different $log(c')$.

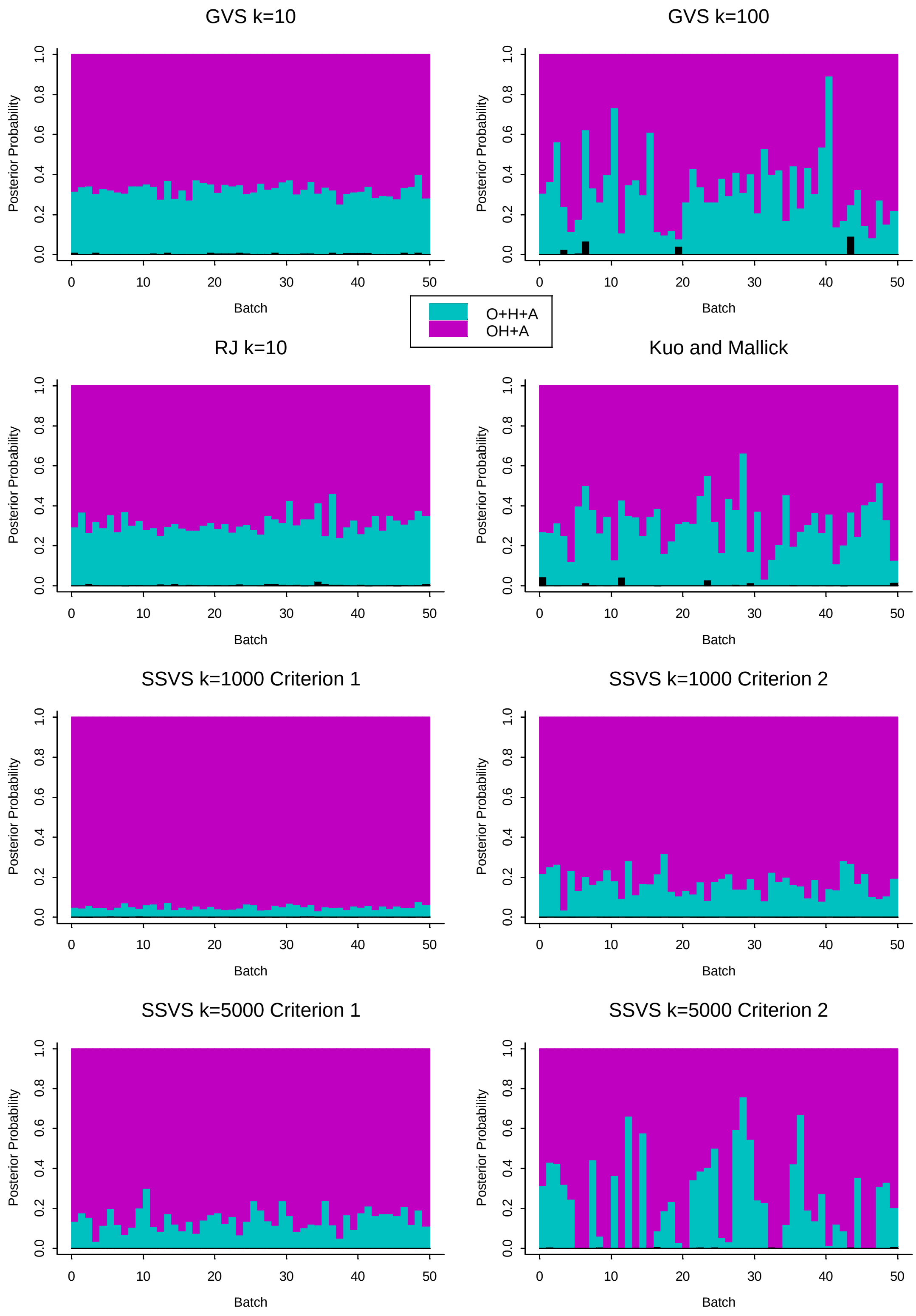
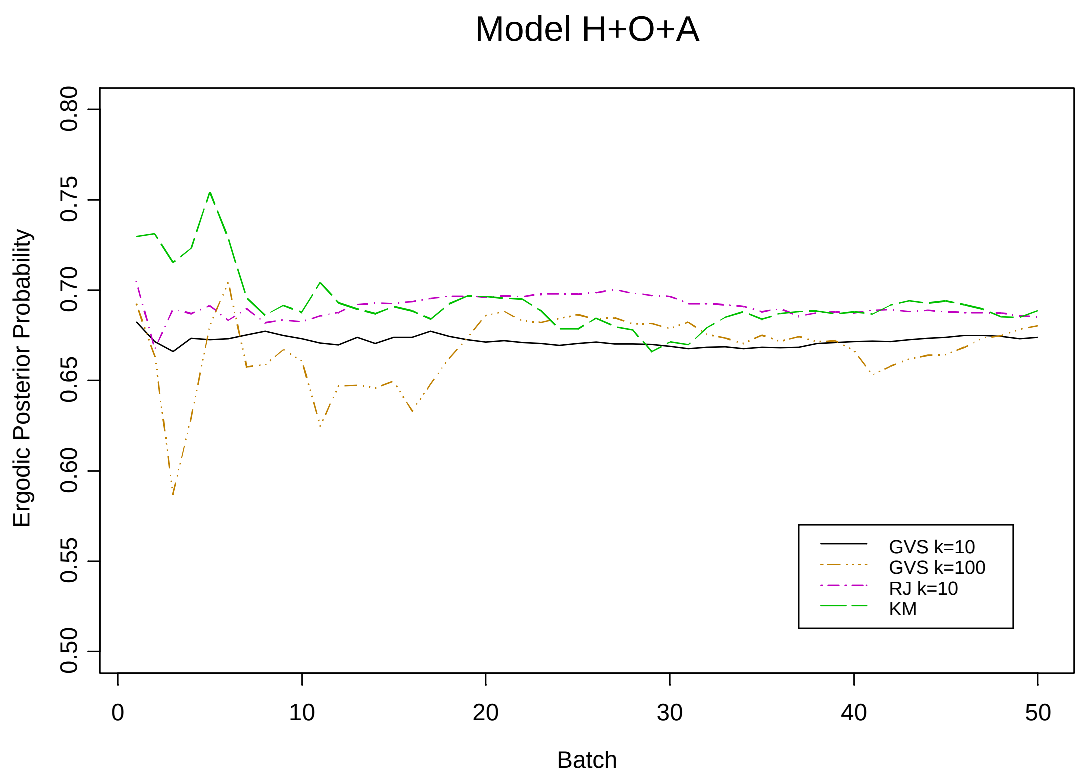


Figure 4.11: $3 \times 2 \times 4$ Contingency Table Example: Barplot Comparison of Different MCMC Model Selection Methods.



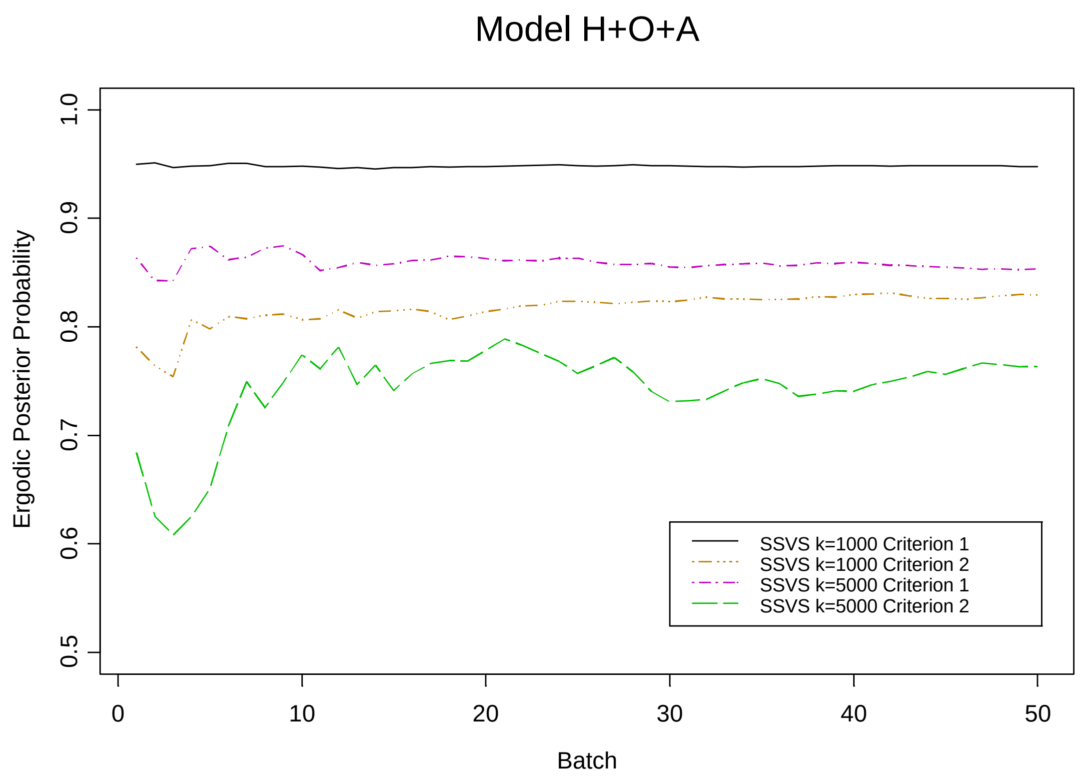


Figure 4.12: $3 \times 2 \times 4$ Contingency Table Example: Ergodic Posterior Probability of Model $H + O + A$ for Different MCMC Model Selection Methods.

4.6.3 Binomial Regression Models

Binomial models are used when the dependent variable is binary and interest lies in determining the factors that affect the resulted probability. The model formulation is given by

$$Y_i \sim Binomial(p_i, N_i).$$

The probability p_i has a different form depending on the link function $g(p_i)$ used. The most frequent link functions are the logit [$p_i = g^{-1}(\eta_i) = exp(\eta_i)/\{1 + exp(\eta_i)\}$] the probit [$p_i = g^{-1}(\eta_i) = \Phi^{-1}(\eta_i)$] and the complementary log-log [$p_i = g^{-1}(\eta_i) = 1 - exp(-e^{\eta_i})$]. Therefore the likelihood is given by

$$f(\boldsymbol{y}|\boldsymbol{\beta}_{(\boldsymbol{\gamma})}, \boldsymbol{\gamma}) = exp\left\{\sum_{i=1}^{n}\left(\begin{array}{c} N_i \\ y_i \end{array}\right) + \sum_{i=1}^{n} y_i log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} N_i log(1-p_i)\right\}.$$

For binomial models the procedure for sampling model parameters is equivalent to the Poisson log-linear models. The general sampling procedure is given by

1. Generate $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ from

$$f(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma}, \boldsymbol{y}) \propto exp\left\{\sum_{i=1}^{n} y_i log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} N_i log(1-p_i)\right\} f\left(\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\boldsymbol{\gamma}\right)$$

   substituting p_i by $g^*\left(\left[\boldsymbol{X}_{(\boldsymbol{\gamma})}\boldsymbol{\beta}_{(\boldsymbol{\gamma})}\right]_i\right)$ in reversible jump or Carlin and Chib samplers, $\sum_{j\in\mathcal{V}} \gamma_j \boldsymbol{X}_{ij}\boldsymbol{\beta}_j$ in Gibbs variable selection and $g^*\left([\boldsymbol{X\beta}]_i\right)$ for SSVS. When the canonical (logit) link is used the above full conditional posteriors (without their normalising constants) may be given by (4.22) in Carlin and Chib variants, (4.23) in Gibbs variable selection and (4.24) in SSVS:

$$exp\left\{\sum_{i=1}^{n} y_i log\left[\boldsymbol{X}_{(\gamma)}\beta_{(\gamma)}\right]_i - \sum_{i=1}^{n} N_i log\left[1 + exp\left(\left[\boldsymbol{X}_{(\gamma)}\beta_{(\gamma)}\right]_i\right)\right]\right\} f(\beta_{(\boldsymbol{\gamma})}|\gamma) \quad (4.22)$$

$$exp\left\{\sum_{i=1}^{n} y_i log\left(\sum_{j\in\mathcal{V}} \gamma_j \boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right) - \sum_{i=1}^{n} N_i log\left[1 + exp\left(\sum_{j\in\mathcal{V}} \gamma_j \boldsymbol{X}_{ij}\boldsymbol{\beta}_j\right)\right]\right\} f(\beta_{(\boldsymbol{\gamma})}|\gamma) \quad (4.23)$$

$$exp\left\{\sum_{i=1}^{n} y_i log\left[\boldsymbol{X}\beta\right]_i - \sum_{i=1}^{n} N_i log\left[1 + exp\left([\boldsymbol{X}\beta]_i\right)\right]\right\} f\left(\beta|\gamma\right) \quad (4.24)$$

   where $\boldsymbol{X}_{ij}$ is the $1 \times d_j$ sub-matrix corresponding to i observation and j term.

2. The scale parameter is known and therefore we do not have to estimate it.

Further details are omitted and we directly present an illustrated example.

4.6.3.1 A Logistic Regression Example

We consider a dataset analysed by Healy (1988). The data, presented in Table 4.11, reflect the relationship between the number of survivals, the patient condition (more or less severe) and the received treatment (antitoxin medication or not). We wish to select one of the five

	Antitoxin	Death	Survivals
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

Table 4.11: Logistic Regression Example: Healy (1988) Dataset.

possible nested logistic regression models with response variable the number of survivals and explanatory factors the patient condition and the received treatment. The full model is given by

$$Y_{il} \sim Bin(N_{il}, p_{il}), \quad log\left(\frac{p_{il}}{1-p_{il}}\right) = \mu + a_i + b_l + (ab)_{il}, \quad i, l = 1, 2,$$

where Y_{il}, N_{il} and p_{il} are the number of survivals, the total number of patients and the probability of survival under i level of severity and l treatment respectively; μ, a_i, b_l and $(ab)_{il}$ are the model parameters corresponding to the constant term, i level of severity, l level of treatment, and interaction of i severity and l treatment.

We consider, for our illustration, the reversible jump, the Metropolised version of Carlin and Chib's method presented in Section 4.3.1, the Gibbs variable selection introduced in Section 4.1.1, the Kuo and Mallick (1998) method presented in Section 3.4.2 and the SSVS method presented in 3.4.1. A rough guideline for our comparisons is an approximation to Bayes factor B_{10} of model m_1 against model m_0 based on Bayes information criterion (BIC) given by (2.4) with sample size given by the sum of all Bernoulli trials, that is $\sum_{il} N_{il}$ for this logistic regression example; see Raftery (1996a) for details. Calculation of all Bayes factors against the full model leads immediately to posterior model probabilities.

Assuming the usual sum-to-zero constraints, the parameter vector for the full model is given by $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3) = (\mu, a_2, b_2, (ab)_{22})$. We use the same $N(0, \Sigma)$ prior distribution for each β_j, $j = 0, \ldots, 3$, under all five models. Following the ideas of Dellaportas and Forster (1999) we choose $\Sigma = 4 \times 2$ as a variance which gives a diffuse but proper prior. In SSVS we used as variance divisor $k_j = 1000$ for all model parameters in order to give approximately the same results as the other methods.

All Markov chains were initiated at the full model with starting points $\beta_j = 0$ for all $j = 0, \ldots, 3$. For the reversible jump and Metropolised Carlin and Chib methods we always propose only a 'neighbouring' model which differs from the current model by one term, hence $j(m, m) = 0$. The proposal distributions for β_j in the above methods as well as the pseudopriors in Gibbs variable selection are $N(\bar{\mu}_j, S_j)$ where $\bar{\mu}_j$ and S_j were estimated from a pilot run of 500 iterations in the full model, after discarding the first 100 as burn-in iterations. The resulting values of $\bar{\mu}_j$ and S_j turned out to be $(-0.47, -0.87, 0.56, -0.17)$ and $(0.27, 0.27, 0.28, 0.27)$ for $j = 0, \ldots, 3$ respectively. Within each model, updating of the parameters β_j was performed via Gibbs sampling steps as described in Dellaportas and Smith (1993). Finally, each Markov chain ran for 110,000 iterations and the output summaries are based on ergodic averages taken on the last 100,000 iterations. All of the MCMC approaches took a similar length of time (around 4 minutes in a Pentium 100 PC). The full results are given in Table 4.12.

Model	Deviance	AP	SSVS	GVS	KM	RJ	MCC
Constant	18.656	0.004	0.011	0.005	0.005	0.005	0.005
A	4.748	0.460	0.498	0.493	0.496	0.491	0.491
B	12.171	0.011	0.017	0.011	0.010	0.012	0.012
$A+B$	0.368	0.462	0.416	0.439	0.436	0.439	0.440
AB	0.000	0.063	0.057	0.051	0.053	0.052	0.052

Table 4.12: Logistic Regression Example: Posterior Model Probabilities; AP = approximate probabilities, SSVS = stochastic search variable selection, GVS = Gibbs variable selection, KM = Kuo and Mallick Gibbs sampler, RJ = reversible jump, MCC = Metropolised Carlin and Chib method.

As would be hoped, all MCMC methods give similar results, with the combined probability of the two most probable models to be at least 0.93. SSVS gives slightly different results

	Batch Standard Deviation	
	Model A	Model $A+B$
GVS	0.012	0.010
KM	0.021	0.019
RJ	0.017	0.014
MCC	0.016	0.013
SSVS	0.196	0.168

Table 4.13: Logistic Regression Example: Batch Standard Deviation of Posterior Model Probabilities.

as expected. The resulting 110,000 iterations were divided in 44 batches of length of 2,500 iterations. Probabilities of the best two models over different batches are given in Figure 4.14 while ergodic probabilities are given in Figure 4.15. Assessment of convergence can be done via these figures and Table 4.13 gives the between batches standard deviation. The smaller the standard deviation the quicker the convergence. From both plots and standard deviations we clearly see that SSVS demonstrates more variability than all the other methods and therefore we need more iterations to reach convergence. The smallest standard deviation is achieved by Gibbs variable selection while standard deviation of Kuo and Mallick sampler is about twice as much. Reversible jump and Metropolised Carlin and Chib have about the same standard deviation.

We finally performed a robustness analysis. We assume that $c^2 = 4 \times c'$ and we calculated the posterior probabilities for various values of c'. For values of c' between 2 and 100, the model including the severity effect (A) is supported. Its probability reaches its maximum (0.834) at the value $c' = 57.6$ (estimated by the fitted model). Note that the more complicated model which includes both severity and antitoxin effect decreases its probability when the prior parameter c' increases.

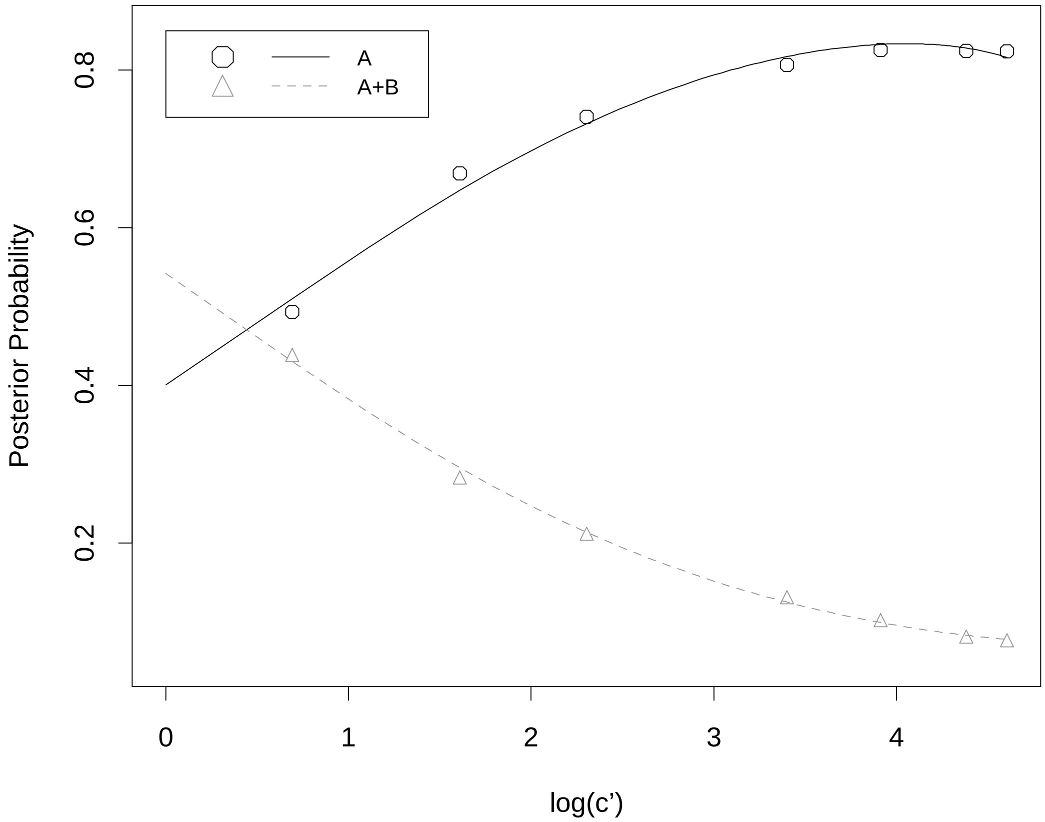


Figure 4.13: Logistic Regression Example: Posterior Probabilities for Different Prior Distributions.

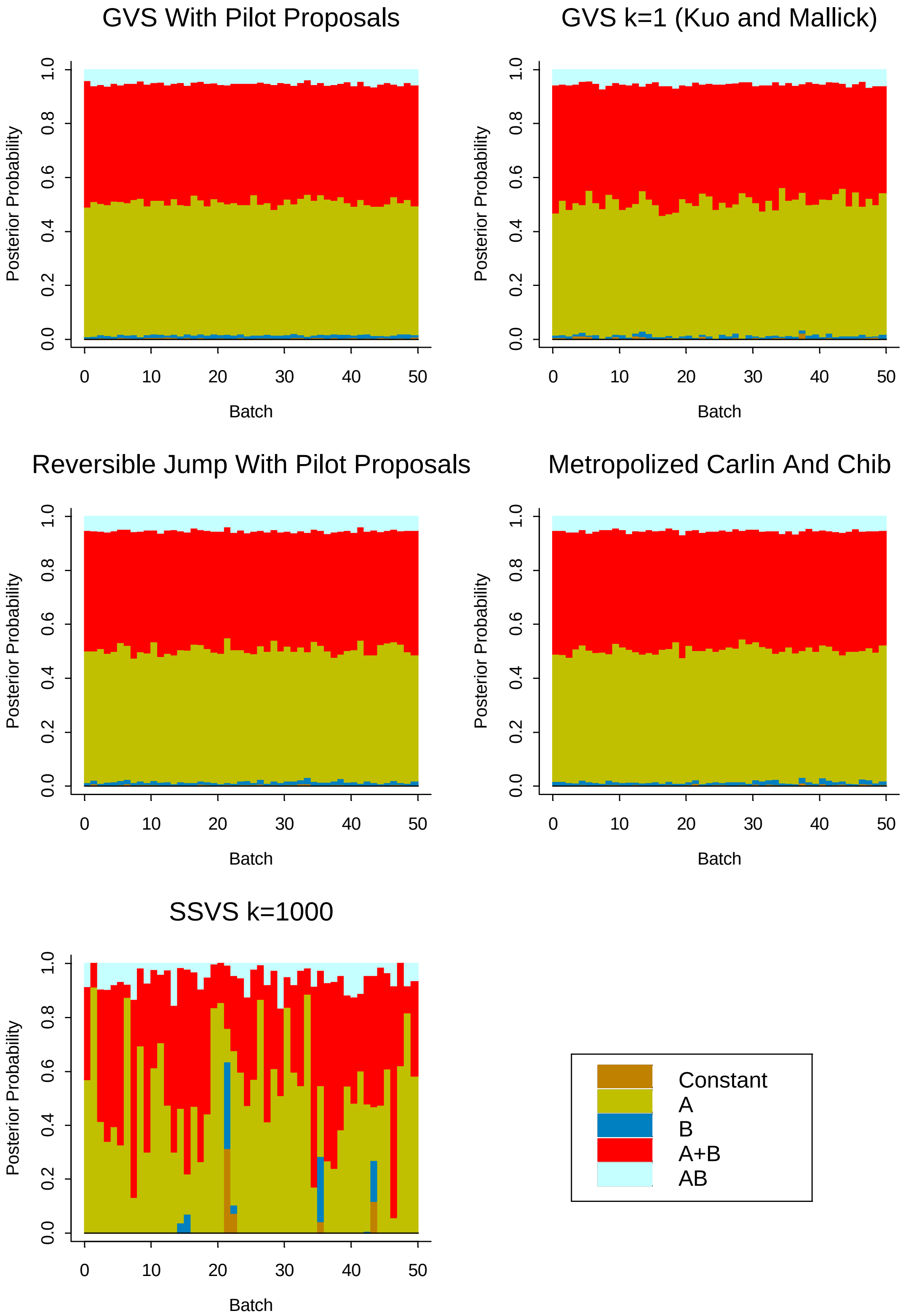


Figure 4.14: Logistic Regression Example: Batch Posterior Probabilities.

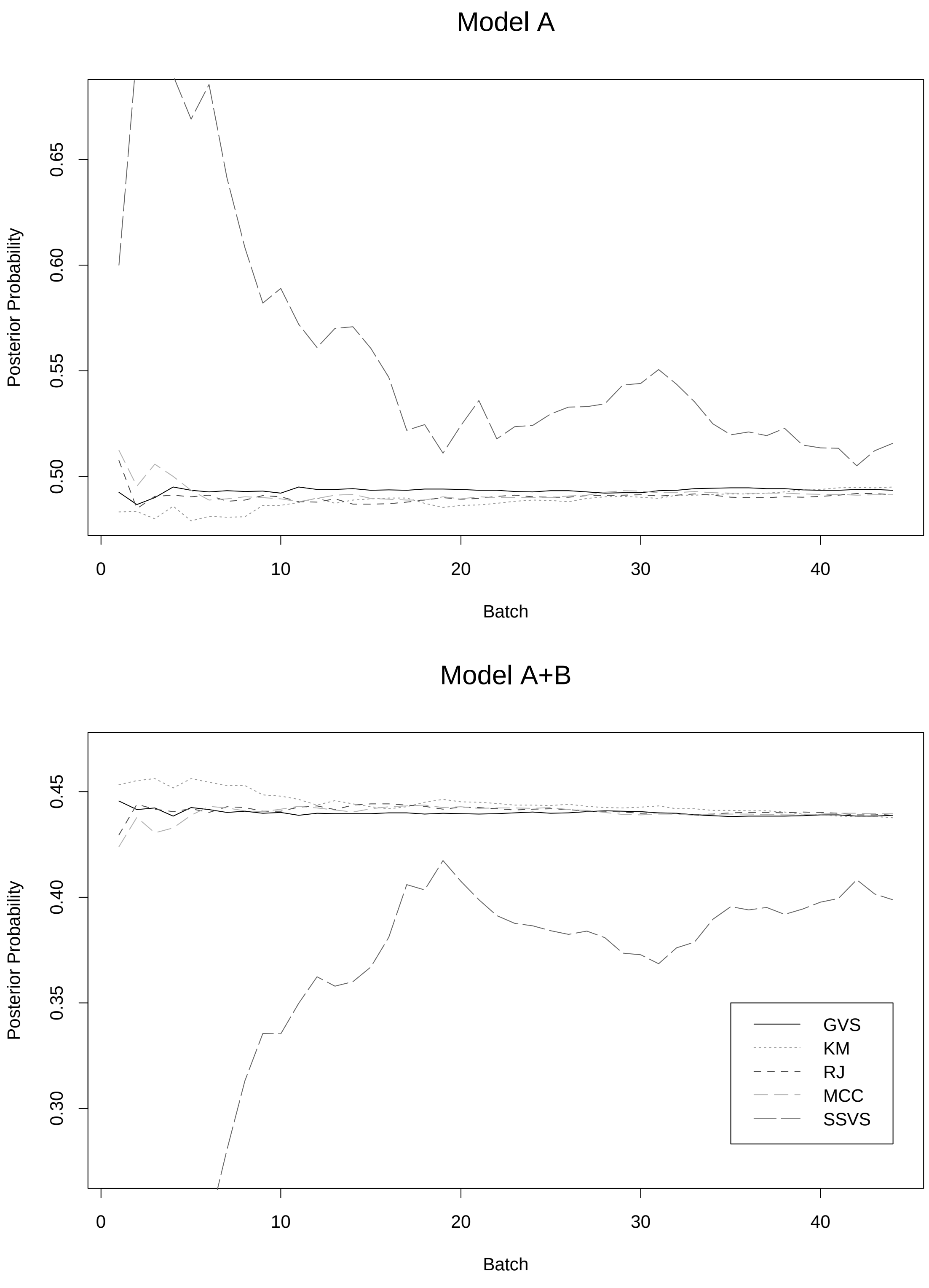


Figure 4.15: Logistic Regression Example: Ergodic Posterior Probabilities.

Chapter 5

Simultaneous Covariate and Structural Identification in Generalised Linear Models

We concentrate in model selection aspects that may appear in generalised linear models. The usual latent model indicator $m \in \mathcal{M}$ is used to represent the general model formulation including error structure, link function and covariate selection. The set of all possible models $\mathcal{M}$ can be written as a product of two subsets $\mathcal{S} \times \{0,1\}^p$, where $\mathcal{S}$ is the set of all structural properties and $\{0,1\}^p$ the set of all possible combinations of covariates included in the model. Therefore, there is a one-to-one transformation $G : \mathcal{M} \to \mathcal{S} \times \{0,1\}^p$. The set $\mathcal{S}$ can be further analysed to a product of various other sets such as the set of available links $\mathcal{L}$ or the set of error distributions $\mathcal{D}$. Covariates to be selected are denoted by the usual vector of binary indicator parameters $\boldsymbol{\gamma}$ ($\gamma_j = 1$ indicates that the j term is included in the model while $\gamma_j = 0$ indicates that the j term is excluded from the model). For example, in a model selection problem where we account uncertainty on the error distribution, link function and covariate selection we use the latent variables $(D, L, \boldsymbol{\gamma}) \in \mathcal{D} \times \mathcal{L} \times \{0,1\}^p$.

In generalised linear models interest usually lies in covariate identification. However, the selection of structural properties is closely related with the validity of the model, its sensitivity to extreme or outlying values, and the structure of randomness in the response variable that we study. Therefore here we discuss and present some implementation details when we are interested in other modelling aspects. We focus in link selection but the methodology developed for the link identification can be implemented without any further problems to the determination of other structural properties. The main theory and results of this chapter have also given in a form of a technical report; see Ntzoufras *et al.* (1999b)

The following chapter is organized into five sections. The first one gives variable and link selection details. Detailed samplers are presented together with a straightforward technique to develop 'equivalent' (in a loose sense) prior distributions among different link functions. The second section introduces a method for outlier identification while the third discusses and compares link and transformation selection. The fourth section discusses the problem of distribution selection and gives details of special cases. The final section presents an illustrative example using simultaneous variable and link selection methods.

5.1 Covariate and Link Function Identification

Variable and link selection breaks into two steps by using the latent variables $\boldsymbol{\gamma}$ for covariate selection and L for link identification. Then we may use any of the available MCMC methods for the covariate selection. We may also incorporate both link and variable selection in one reversible jump Metropolis step though it is quite inefficient. The linear predictor will slightly change to

$$g_L(\mu_i) = \eta_i^L, \quad \boldsymbol{\eta}^L = \sum_{j \in \mathcal{V}} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_{j,(L)} \tag{5.1}$$

where $\boldsymbol{\beta}_{j,(L)}$ is the parameter vector that corresponds to j term and L link while the vector of the full model for L link, is noted by $\boldsymbol{\beta}_{(L)}$.

5.1.1 'Equivalent' Priors for Non-canonical Link Functions

An important issue in link selection is the specification of prior distributions that represent the 'equivalent' beliefs for different link functions. Suppose that we use independent prior distributions for each term conditionally on the link function used. Then the prior of type

$$\boldsymbol{\beta}_{j,(L)} | \gamma_j = 1, L \sim N(\mathbf{0}, \boldsymbol{\Sigma}_j^L), \text{ for all } L \in \mathcal{L}$$

can be utilized.

A simple approach is based on the notion of semiautomatic selection used in SSVS by George and McCulloch (1993). In the simple case of one dimensional regressors we simply use the equation $[Var(\hat{\beta}_j^L)]^{-1}\Sigma_j^L = [Var(\hat{\beta}_j^C)]^{-1}\Sigma_j^C$ resulting in

$$\Sigma_j^L = \frac{Var(\hat{\beta}_j^L)}{Var(\hat{\beta}_j^C)}\Sigma_j^C$$

where Σ_j^C and Σ_j^L are the prior variances for the canonical link function C and any other link function $L \in \mathcal{L}$, while $Var(\hat{\beta}_j^C)$ and $Var(\hat{\beta}_j^L)$ are the standard errors of the maximum likelihood estimates $\hat{\beta}_j^C$ and $\hat{\beta}_j^L$ of links C and L, respectively.

In more complicated cases we may use multivariate normal distributions and adjust the variance of each parameter and the prior covariances of parameters β_k and β_l using the equation

$$Cov(\beta_k^L, \beta_l^L) = \frac{Cov(\hat{\beta}_k^L, \hat{\beta}_l^L)}{Cov(\hat{\beta}_k^C, \hat{\beta}_l^C)}Cov(\beta_k^C, \beta_l^C).$$

A more sophisticated adjustment can be based on the first order approximation of Taylor expansion. This is given by

$$\eta^L = g_L(\mu) \approx g_L(\mu_0) + (\mu - \mu_0)g_L'(\mu_0) \tag{5.2}$$

where $L \in \mathcal{L}$ is a link indicator $g_L(\mu)$ is the link function and η^L the linear predictor corresponding to L indicator. The linear function

$$TE_L(\mu_0) = [g_L(\mu_0) - \mu_0 g_L'(\mu_0)] + \mu g_L'(\mu_0)$$

will be called Taylor expansion linear approximation of link function g_L around μ_0. Then, solving the equation

$$\frac{\eta^{L_1} - g_L(\mu_0)}{g_{L_1}'(\mu_0)} = \frac{\eta^{L_2} - g_L(\mu_0)}{g_{L_2}'(\mu_0)} \tag{5.3}$$

we have

$$\eta^{L_1} = \delta^*_{L_1L_2}(\mu_0)\eta^{L_2} + \delta^{**}_{L_1L_2}(\mu_0) \tag{5.4}$$

$$\delta^*_{L_1L_2}(\mu_0) = \frac{g_{L_1}'(\mu_0)}{g_{L_2}'(\mu_0)} \tag{5.5}$$

$$\delta^{**}_{L_1L_2}(\mu_0) = g_{L_1}(\mu_0) - \delta^*_{L_1L_2}(\mu_0)\eta^{L_2}g_{L_2}(\mu_0). \tag{5.6}$$

The most frequent application of link selection in generalised linear models is considered in binomial models where the most popular links are logistic, probit and complementary log-log. A common choice for μ_0 is $\mu_0 = p_0 = 0.5$ but this approximation is not effective when data contain many values resulting in binomial probabilities close to one or zero. In

L	Link	$g(p_0)$	$g'(p_0)$
L_1	Logit	$log[p_0/(1-p_0)]$	$[p_0(1-p_0)]^{-1}$
L_2	Probit	$\Phi^{-1}(p_0)$	$[\varphi(p_0)]^{-1}$
L_3	clog-log	$log[-log(1-p)]$	$-[(1-p_0)log(1-p_0)]^{-1}$
L_4	log-log	$log[-log(p)]$	$[p_0 log(p_0)]^{-1}$

Table 5.1: Table of Taylor Expansion for Binomial Example.

L	Link	$g(0.5)$	$g'(0.5)$	TE(0.5)	$\delta^*_{LL_1}$	$\delta^{**}_{LL_1}$
L_1	Logit	0.000	4.000	$4p-2$	1.000	0.000
L_2	Probit	0.000	2.507	$2.507p-1.253$	0.627	0.000
L_3	clog-log	-0.367	2.885	$2.885p-1.809$	0.721	-0.367
L_4	log-log	-0.367	-2.885	$-2.885p+1.076$	-0.721	-0.367

Table 5.2: Table of Taylor Expansion (TE) for Binomial Example; $p_0 = 0.5$.

such cases, a more effective choice is provided by the mean or median probability of available binomial trials. Details for the binomial case are given in Tables 5.1 and 5.2.

We suggest that the linear transformation

$$\beta_{0,(L)} = \delta^*_{LL_1}(\mu_0)\beta_{0,(L_1)} + \delta^{**}_{LL_1}(\mu_0) \tag{5.7}$$

$$\boldsymbol{\beta}_j^L = \delta^*_{LL_1}(\mu_0)\boldsymbol{\beta}_j^{L_1} \tag{5.8}$$

is a good approximation between the parameters of the two different links since they satisfy (5.4).

An alternative approach given by Raftery (1996a) is described in Section 3.2.1.3. We may also use the prior (3.10) with covariance matrix proportional to the inverse of the observed information matrix as defined by (3.13) or the unit root prior approach. Similar approaches can be used for transformation identification or distribution selection problems.

5.1.2 Reversible Jump Link Selection for Given Covariate Structure

The Metropolis step for link selection involves switching link functions while the dimension of the two models remain unchanged and equal to $d(\gamma)$. Two types of reversible jump are appropriate for link selection. The first is to use a suitable transformation without generating any additional terms. The identity transformation may also be used but in some cases (where posterior distribution of parameters in different links are quite distinct) it turned out to be very inefficient. The second reversible jump type involves a Metropolised Carlin and Chib step where all the parameters are generated from appropriate proposals.

The reversible jump link step with transformation has acceptance probability

$$\alpha = min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{(\gamma,L')}, \gamma, L')f(\boldsymbol{\beta}'_{(\gamma,L')}|\gamma, L')f(\gamma, L')j(L', L)}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(\gamma,L)}, \gamma, L)f(\boldsymbol{\beta}_{(\gamma,L)}|\gamma, L)f(\gamma, L)j(L, L')}\left|\frac{\partial h_{L,L'}(\boldsymbol{\beta}_{(\gamma,L)})}{\partial(\boldsymbol{\beta}_{(\gamma,L)})}\right|\right) \tag{5.9}$$

with $h_{L,L'}(\boldsymbol{\beta}_{(\gamma,L)}) = \boldsymbol{\beta}'_{(\gamma,L')}$.

A simple but well performed transformation can be achieved by using the first two terms of the Taylor expansion given by (5.2). The simple linear transformation given by equations (5.7) and (5.8) can be used to assist the chain to jump easily from one link to another. This transformation has the property that the expected values μ_i remain approximately equal among different links since

$$\begin{aligned}
\boldsymbol{\eta}^{L'} &= \beta_{0,(L')} + \sum_{j\in\mathcal{V}\setminus\{0\}} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_{j,(L')} \\
&= \delta^{*}_{L'L}(\mu_0)\beta_{0,(L)} + \delta^{**}_{L'L}(\mu_0) + \delta^{*}_{L'L}(\mu_0) \sum_{j\in\mathcal{V}\setminus\{0\}} \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_{j,(L)} \\
&= \delta^{*}_{L'L}(\mu_0)\boldsymbol{\eta}^{L} + \delta^{**}_{L'L}(\mu_0),
\end{aligned}$$

which is the result we get when we equate the Taylor approximated expected values for two different links; see (5.3) and (5.4).

In general, we may use the value of $p_0 = 0.5$ but in some situations may not be efficient (for example when all the probabilities are small). A better choice for p_0 may be the average probability of the binomial trials. This linear transformation improves the reversible jump step efficiently without having to calculate a complicated Jacobian but the simple $|\delta^{*d(\gamma)}_{LL'}|$.

An alternative approach is to consider Metropolised Carlin and Chib and generate all model parameters of the new proposed link from an appropriate proposal distribution. In such case the Metropolis step is given by

$$\alpha = min\left(1, \frac{f(\boldsymbol{y}|\boldsymbol{\beta}'_{(\boldsymbol{\gamma},L')}, \boldsymbol{\gamma}, L')f(\boldsymbol{\beta}'_{(\boldsymbol{\gamma},L')}|\boldsymbol{\gamma}, L')f(\boldsymbol{\gamma}, L')j(L', L)q_L(\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)})}{f(\boldsymbol{y}|\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}, \boldsymbol{\gamma}, L)f(\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}|\boldsymbol{\gamma}, L)f(\boldsymbol{\gamma}, L)j(L, L')q_{L'}(\boldsymbol{\beta}'_{(\boldsymbol{\gamma},L')})}\right) \tag{5.10}$$

Note that the Jacobian here is equal to one since $h_{m,m'}(\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}, \boldsymbol{\beta}'_{(\boldsymbol{\gamma},L')}) = (\boldsymbol{\beta}'_{(\boldsymbol{\gamma},L')}, \boldsymbol{\beta}_{(\boldsymbol{\gamma},L)})$. The proposal distributions can have the form $q_L(\boldsymbol{\beta}_{(\boldsymbol{\gamma},L)}) = N(\bar{\boldsymbol{\mu}}_{\beta_{(\gamma,L)}}, \boldsymbol{S}_{(\boldsymbol{\gamma},L)})$, where $\bar{\boldsymbol{\mu}}_{\beta_{(\gamma,L)}}$ and $\boldsymbol{S}_{(\boldsymbol{\gamma},L)}$ can be taken by a pilot chain for the corresponding model.

5.1.3 Gibbs Variable and Link Selection for Generalised Linear Models

This section demonstrates how we can extend Gibbs variable selection methods to other selection problems such as link selection. Similar methodology may be followed for distribution selection problems.

The first approach in applying the variable and link selection via Gibbs is to generalise Gibbs variable selection or SSVS. We use again the same notation as in Gibbs variable selection and therefore $\boldsymbol{\beta}$ denotes the parameter vector of the full model for any link. The resulting posterior conditional distributions for Gibbs variable selection are equivalent to (4.1), (4.2) and (4.3) adding the link indicator L while the conditional posterior for the link selection step is given by

$$f(L|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, L)f(\boldsymbol{\beta}|\boldsymbol{\gamma}, L)f(\boldsymbol{\gamma}, L)}{\sum\limits_{L' \in \mathcal{L}} f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, L')f(\boldsymbol{\beta}|\boldsymbol{\gamma}, L')f(\boldsymbol{\gamma}, L')} \tag{5.11}$$

where $f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, L)$ is the likelihood for the L link.

The main drawback of the above chain is that when the distance of the posterior distributions of the parameters for different links is large then the part of the prior referring to the coefficients outside the current model will slow down the link jumps. A proposed modification is to use of same pseudoprior for all links resulting to their elimination from the link selection step. This will considerably slow down the variable selection step, since proposed values are not optimal, making the Markov chain inefficient. For these reason we shall develop other Gibbs based samplers for simultaneous variable and link selection.

Similarly SSVS can be extended to stochastic search variable and link selection. This extension will be given by posterior equivalent to (3.24), (3.25) and the above link step (equation 5.11). Stochastic search variable and link selection demonstrates convergence problems similar to Gibbs variable and links selection sampler especially when the parameters of different links are distinct.

A possible solution to the above problem is given by combining Gibbs variable selection for variable selection and Carlin and Chib approach for link selection. Instead of $\boldsymbol{\beta}$ for all link functions we use one vector $\boldsymbol{\beta}_{(L)}$ for each $L \in \mathcal{L}$. Therefore we need to define $f(\boldsymbol{\beta}_{(L)}|\boldsymbol{\gamma}, L')$ for all $L, L' \in \mathcal{L}$. When $L = L'$ then the parameter vector $\boldsymbol{\beta}_{(L)}$ is generated from $f(\boldsymbol{\beta}_{(L)}|\boldsymbol{\gamma}, L, \boldsymbol{y})$ which is equivalent to Gibbs variable selection steps given by (4.1) and (4.2) adding the link indicator L in all conditional distributions and to all model vectors $\boldsymbol{\beta}$, $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ and $\boldsymbol{\beta}_{(\backslash\boldsymbol{\gamma})}$. In the case where $L \neq L'$ we generate $\boldsymbol{\beta}_{(L)}$ from the pseudoprior $f(\boldsymbol{\beta}_{(L)}|\boldsymbol{\gamma}, L' \neq L)$. The link selection step is given by the conditional posterior

$$f(L|\{\boldsymbol{\beta}_{L'},\ L' \in \mathcal{V}\}, \boldsymbol{\gamma}, \boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}_{(L)}, \boldsymbol{\gamma}, L) f(\boldsymbol{\gamma}, L) \prod_{L' \in \mathcal{L}} f(\boldsymbol{\beta}^{(L')}|\boldsymbol{\gamma}, L)}{\sum_{\ell \in \mathcal{L}} f(\boldsymbol{y}|\boldsymbol{\beta}_{(\ell)}, \boldsymbol{\gamma}, \ell) f(\boldsymbol{\gamma}, \ell) \prod_{L' \in \mathcal{L}} f(\boldsymbol{\beta}^{(L')}|\boldsymbol{\gamma}, \ell)}. \tag{5.12}$$

The above Gibbs sampler solves the problem appeared in Gibbs variable selection or SSVS extension to link identification problems. The major drawback of this Gibbs sampler is the prerequisite of proposing many parameters from the pseudopriors. This can be solved if we metropolise this step resulting to Metropolised Carlin and Chib link selection step described in the previous section. This sampler was mainly constructed to illustrate how we can create a Gibbs sampler that can easily be implemented by easy-to-use packages such as BUGS. Similar procedures can be used to construct flexible samplers that give accurate results in reasonable time. We can also construct a Gibbs sampler that applies SSVS methodology for variable selection and Carlin and Chib methodology for link selection as described above.

5.1.4 Metropolised Gibbs Sampler for Link Selection

The optimal Metropolised Gibbs sampler described in Section 4.5.1.2 for discrete random variables can also be implemented for the link selection samplers. We develop this variant in which we propose a new link $L' \neq L$ from the distribution

$$j(L, L') = \frac{f(\boldsymbol{y}|\boldsymbol{\beta}_{(L')}, \boldsymbol{\gamma}, L') f(\boldsymbol{\gamma}, L') \prod_{L'' \in \mathcal{L}} f(\boldsymbol{\beta}_{(L'')}|\boldsymbol{\gamma}, L')}{\sum_{\ell \in \mathcal{L}\backslash\{L\}} f(\boldsymbol{y}|\boldsymbol{\beta}_{(\ell)}, \boldsymbol{\gamma}, \ell) f(\boldsymbol{\gamma}, \ell) \prod_{L'' \in \mathcal{L}} f(\boldsymbol{\beta}_{(L'')}|\boldsymbol{\gamma}, \ell)}$$

and accept the proposed term with acceptance probability

$$\alpha = min\left(1, \frac{\sum_{\ell \in \mathcal{L}\setminus\{L\}} f(\boldsymbol{y}|\boldsymbol{\beta}_{(\ell)}, \gamma, \ell) f(\gamma, \ell) \prod_{L'' \in \mathcal{L}} f(\boldsymbol{\beta}_{(L'')}|\gamma, \ell)}{\sum_{\ell \in \mathcal{L}\setminus\{L'\}} f(\boldsymbol{y}|\boldsymbol{\beta}_{(\ell)}, \gamma, \ell) f(\gamma, \ell) \prod_{L'' \in \mathcal{L}} f(\boldsymbol{\beta}_{(L'')}|\gamma, \ell)}\right).$$

Although the above sampler seems complicated it does not require more computations than Gibbs variable and Carlin and Chib link selection defined above. On the other hand, it is more complicated than both versions of reversible jump for link selection proposed in Section 5.1.2. The above sampler is given in a simpler form if we set

$$\Xi(\boldsymbol{\beta}^*, \gamma, \ell) = f(\boldsymbol{y}|\boldsymbol{\beta}_{(\ell)}, \gamma, \ell) f(\gamma, \ell) \prod_{L'' \in \mathcal{L}} f(\boldsymbol{\beta}_{(L'')}|\gamma, \ell),$$

$$\Xi(\boldsymbol{\beta}^*, \gamma) = \sum_{\ell \in \mathcal{L}} \Xi(\boldsymbol{\beta}_{(\ell)}, \gamma, \ell),$$

where $\boldsymbol{\beta}^* = \{\boldsymbol{\beta}_{(L'')},\ L'' \in \mathcal{L}\}$. The proposal is now given by

$$j(L, L') = \frac{\Xi(\boldsymbol{\beta}^*, \gamma, L')}{\Xi(\boldsymbol{\beta}^*, \gamma) - \Xi(\boldsymbol{\beta}^*, \gamma, L)}$$

and accept the proposed term with probability

$$\alpha = min\left(1, \frac{\Xi(\boldsymbol{\beta}^*, \gamma) - \Xi(\boldsymbol{\beta}^*, \gamma, L)}{\Xi(\boldsymbol{\beta}^*, \gamma) - \Xi(\boldsymbol{\beta}^*, \gamma, L')}\right).$$

5.1.5 Other Approaches in Link Selection

The above link selection methods considered a limited number of distinct link functions. More general link functions can be adopted by considering a family of link function depending on one or more (unknown) parameters. The set of all possible links will depend on the possible values that these parameters can take. For example Mallick and Gelfand (1994) consider mixtures of beta distributions while Basu and Mukhopadhyay (1994) normal scale mixtures. Lang (1997) considers the link function

$$g_\varrho(\mu_i) = m_1(\varrho) F_{-\infty}(\mu_i) + m_2(\varrho) F(\mu_i) + m_3(\varrho) F_\infty(\mu_i)$$

where ϱ is a mixing parameter to be estimated and $m_i(\varrho)$ for $i = 1, 2, 3$ are mixing functions. As mixing functions, he proposed $m_1(\varrho) = exp(-e^{3.5\varrho+2})$ $m_3(\varrho) = exp(-e^{-3.5\varrho+2})$ and $m_2(\varrho) = 1 - m_1(\varrho) - m_3(\varrho)$ while $F_{-\infty}(\mu) = 1 - exp(-e^{\mu})$ (extreme minimum value function), $F_\infty(\mu) = exp(-e^{\mu})$ (extreme maximum value function) and $F(\mu) = e^{\mu}/(1 + e^{\mu})$

(logistic function). For the mixing parameter ϱ, a normal prior distribution was suggested. A straightforward modification can be done by

$$g_{\varrho}(\mu) = m_1 F_{-\infty}(\mu) + m_2 F(\mu) + (1 - m_1 - m_2) F_{\infty}(\mu)$$

where m_i are now mixing proportions to be estimated. A Dirichlet prior on $[m_1, m_2]$ can be used.

Finally, Albert and Chib (1997) propose to use the family of symmetric links given by

$$\mu_i = g_{\varrho}^{-1}(\eta_i) = \frac{2}{\varrho} \times \frac{\eta_i^{\varrho} - (1-\eta_i)^{\varrho}}{\eta_i^{\varrho} + (1-\eta_i)^{\varrho}}$$

where $\varrho = 0.0, 0.4, 1.0$ correspond to the logit, (approximately) probit and linear link respectively. Alternatively for binomial models we may use

$$g_{\varrho}(p_i) = a_{\varrho} \left(\frac{p_i}{1-p_i} \right)^{\varrho} + b_{\varrho}.$$

The choice of $\varrho = 0$, $a_{\varrho} = 1$ and $b_{\varrho} = 0$ corresponds to the logit link. We decided not to proceed in comparative studies since these approaches are of different notion.

5.2 Alternative Procedures for Outlier Identification

An alternative methodology for outlier identification is proposed in this subsection. This proposed method removes all outliers from the estimation procedure. We also use the latent vector of binary variables $\boldsymbol{v}$ which indicates outliers by $v_i = 0$.

We consider the model

$$Y_i \sim N\left(v_i \sum_{j \in \mathcal{V}} \gamma_j \boldsymbol{X}_{ij} \boldsymbol{\beta}_j + (1 - v_i) o_i, \sigma^2 \right)$$

where o_i is a parameter estimating the expected value of i observation when it is outlier.

Using MC^3 for both variable and outlier identification results in two Metropolis steps. In the first we propose a new model with covariates given by $\boldsymbol{\gamma}'$ differing by $\boldsymbol{\gamma}$ only in one term and accept the move with probability given by (3.39). Then we propose with probability $j(\boldsymbol{v}, \boldsymbol{v}')$ to move from $\boldsymbol{v}$ to $\boldsymbol{v}'$ that differ only i coordinator and accept the move with probability (3.40).

If we adopt priors of the general form

$$\boldsymbol{\beta}_{(\boldsymbol{\gamma},\boldsymbol{v})}|\sigma^2,\gamma,\boldsymbol{v} \sim N\left(\boldsymbol{\mu}_{\beta_{(\gamma,v)}}, \boldsymbol{\Sigma}_{(\boldsymbol{\gamma},\boldsymbol{v})}\sigma^2\right)$$

for model parameters and the prior distribution

$$o_i \sim N(0, K^2\sigma^2)$$

for outlying parameters o_i, then the full conditional posterior distributions are given by a Bernoulli distribution with success probability $O_j/(1+O_j)$ with O_j given by (3.31). In the whole procedure the posterior covariance matrix is now given by

$$\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma},\boldsymbol{v})} = \left(\boldsymbol{X}^T_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{X}_{(\boldsymbol{\gamma},\boldsymbol{v})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma},\boldsymbol{v})}\right)^{-1}$$

where $\boldsymbol{X}_{(\boldsymbol{\gamma},\boldsymbol{v})}$ is the design matrix constructed from non-outlier observations ($v_i = 1$) and from the included covariates ($\gamma_j = 1$). Moreover, the posterior residual sum of squares $SS_{\boldsymbol{\gamma}}$ are substituted by

$$\begin{aligned} SS_{\boldsymbol{\gamma},\boldsymbol{v}} &= \boldsymbol{y}^T_{(\boldsymbol{v})}\boldsymbol{y}_{(\boldsymbol{v})} + (K^2+1)^{-1}\boldsymbol{y}^T_{(\backslash\boldsymbol{v})}\boldsymbol{y}_{(\backslash\boldsymbol{v})} + \boldsymbol{\mu}^T_{\beta_{(\gamma,v)}}\boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{\mu}_{\beta_{(\gamma,v)}} \\ &\quad -(\boldsymbol{X}^T_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{y}_{(\boldsymbol{v})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{\mu}_{\beta_{(\gamma,v)}})^T\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma},\boldsymbol{v})}(\boldsymbol{X}^T_{(\boldsymbol{\gamma},\,\boldsymbol{v})}\boldsymbol{y}_{(\boldsymbol{v})} + \boldsymbol{\Sigma}^{-1}_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{\mu}_{\beta_{(\gamma,v)}}). \end{aligned}$$

where $\boldsymbol{y}_{(\boldsymbol{v})}$ and $\boldsymbol{y}_{(\backslash\boldsymbol{v})}$ are the vectors with the response values that are included or excluded from the model.

Similarly the outlier identification step will involve sequential generations from a similar Bernoulli with success probability $O^*_j/(1+O^*_j)$ step with

$$O^*_i = (K^2+1)^{-1}\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma},v_i=1,\boldsymbol{v}_{\backslash i})}||\boldsymbol{\Sigma}_{(\boldsymbol{\gamma},v_i=0,\boldsymbol{v}_{\backslash i})}|}{|\boldsymbol{\Sigma}_{(\boldsymbol{\gamma},v_i=1,\boldsymbol{v}_{\backslash i})}||\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma},v_i=0,\boldsymbol{v}_{\backslash i})}|}\right)^{1/2}\left(\frac{SS_{(\boldsymbol{\gamma},v_i=1,\boldsymbol{v}_{\backslash i})}+2b_\tau}{SS_{(\boldsymbol{\gamma},v_i=0,\boldsymbol{v}_{\backslash i})}+2b_\tau}\right)^{-n/2-a_\tau}\frac{f(v_i=1,\boldsymbol{v}_{\backslash i})}{f(v_i=0,\boldsymbol{v}_{\backslash i})}.$$

If a prior distribution equivalent to Smith and Kohn (1996) prior given by

$$\boldsymbol{\beta}_{(\boldsymbol{\gamma},\boldsymbol{v})}|\sigma^2,\gamma,\boldsymbol{v} \sim N\left(\mathbf{0}, c^2\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{X}_{(\boldsymbol{\gamma},\boldsymbol{v})}\right)^{-1}\sigma^2\right)$$

is used, then the ratios of determinants simplify to

$$\left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma},v_i=1,\boldsymbol{v}_{\backslash i})}||\boldsymbol{\Sigma}_{(\boldsymbol{\gamma},v_i=0,\boldsymbol{v}_{\backslash i})}|}{|\boldsymbol{\Sigma}_{(\boldsymbol{\gamma},v_i=1,\boldsymbol{v}_{\backslash i})}||\tilde{\boldsymbol{\Sigma}}_{(\boldsymbol{\gamma},v_i=0,\boldsymbol{v}_{\backslash i})}|}\right)^{1/2} = (c^2+1)^{-d_j/2}$$

and the posterior sum of squares simplifies to

$$SS_{\boldsymbol{\gamma},\boldsymbol{v}} = \boldsymbol{y}^T_{(\boldsymbol{v})}\boldsymbol{y}_{(\boldsymbol{v})} + (K^2+1)^{-1}\boldsymbol{y}^T_{(\backslash\boldsymbol{v})}\boldsymbol{y}_{(\backslash\boldsymbol{v})} - \frac{c^2}{c^2+1}\boldsymbol{y}^T_{(\boldsymbol{v})}\boldsymbol{X}_{(\boldsymbol{\gamma},\boldsymbol{v})}\left(\boldsymbol{X}^T_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{X}_{(\boldsymbol{\gamma},\boldsymbol{v})}\right)^{-1}\boldsymbol{X}^T_{(\boldsymbol{\gamma},\boldsymbol{v})}\boldsymbol{y}_{(\boldsymbol{v})}$$

where $\boldsymbol{y}_{(\boldsymbol{v})}$ and $\boldsymbol{y}_{(\backslash \boldsymbol{v})}$ are the vectors with the response values that are included or excluded from the model. The problem is that this residual variance inherits the Lindley (1957) paradox since now o_i is to be estimated. This problem may be eliminated if we adopt a model formulation in which the model likelihood totally ignores outliers. In this case no additional prior is needed for outlying parameters and hence (K^2+1) and $(K^2+1)^{-1}\boldsymbol{y}^T_{(\backslash \boldsymbol{v})}\boldsymbol{y}_{(\backslash \boldsymbol{v})}$ disappear from the above equations.

Both of the above proposed method can be adopted in Poisson or binomial model or other non-normal models. In such cases advanced MCMC algorithms, such as reversible jump or a Gibbs variable setup, should be adopted in order to identify identify outliers. For example in Poisson model we may adopt the linear predictor

$$\eta_i = log(\lambda_i) = v_i \sum_{j\in\mathcal{V}} \gamma_j \boldsymbol{X}_{ij}\boldsymbol{\beta}_j + (1-v_i)log(o_i)$$

where $o_i|\gamma_i = 0 \sim G(a_o, b_o)$ or in binomial models we may generally consider

$$p_i = \left\{ g^* \left(\sum_{j\in\mathcal{V}} \gamma_j \boldsymbol{X}_{ij}\boldsymbol{\beta}_j \right) \right\}^{v_i} \times o_i^{1-v_i}$$

where $o_i|\gamma_i = 0 \sim Beta(a_o, b_o)$. We additionally need to define proposal distributions for these parameters when they are not included in the likelihood. Straightforward proposal are given by $o_i|\gamma_i = 1 \sim G(y_i, 1)$ in the Poisson case and $o_i|\gamma_i = 1 \sim Beta(y_i, N_i - y_i)$ in binomial.

In the case where we simply ignore outliers from the likelihood, the procedure for sampling model parameters and covariate indicators is the same to the simple covariate selection methods but the summations (or products) involved in the likelihood are limited to non-outlying observations. The additional outlier step is given by a Metropolis or Gibbs step as described in the previous chapter with

$$O_i^* = \frac{f(v_i = 1|\boldsymbol{\beta}_{(\boldsymbol{\gamma}, v_i=1, \boldsymbol{v}_{\backslash i})}, \boldsymbol{\gamma}, \boldsymbol{v}_{\backslash i}, \boldsymbol{y})}{f(v_i = 0|\boldsymbol{\beta}_{(\boldsymbol{\gamma}, v_i=0, \boldsymbol{v}_{\backslash i})}, \boldsymbol{\gamma}, \boldsymbol{v}_{\backslash i}, \boldsymbol{y})}$$

which is again a product of a likelihood ratio, a prior density ratio, a pseudoprior density ratio and prior model probability ratio. If a Gibbs variable selection type algorithm is adopted it is natural to assume that the prior distribution of model parameters and the covariate prior probability are independent of the outlier structure resulting to

$$O_i^* = f(y_i|\boldsymbol{\beta}_{(\boldsymbol{\gamma}, v_i=1, \boldsymbol{v}_{\backslash i})}, \boldsymbol{\gamma}, v_i = 1, \boldsymbol{v}_{\backslash i}) \frac{f(v_i = 1, \boldsymbol{v}_{\backslash i})}{f(v_i = 0, \boldsymbol{v}_{\backslash i})}.$$

The above equation can be interpreted as a measure of how close is observation i to the current model formulation. If the mean value of the current model is close to the mode of the likelihood for i observation then we include this term in the model with high probability.

5.3 Link or Transformation Selection?

On many cases we will face the problem whether we should concentrate on the link or transformation selection. Using transformations of the responses in linear models changes the distribution of the original response while link selection does not affect the distribution of the original response but changes the connection between the mean and the covariates.

For example, if we use the same transforming function then we have $E[g(Y)] = \boldsymbol{X\beta}$ in transformation selection while $g[E(Y)] = \boldsymbol{X\beta}$ in link selection. Using the first two terms of the Taylor expansion we get an approximation for $E[g(Y)]$ in transformation problems resulting to

$$E[g(Y)] \approx g[E(Y)] + g''[E(Y)]Var(Y)/2.$$

$$Var[g(Y)] \approx g'[E(Y)]Var(Y).$$

The above approximation implies that transformation can handle even problems with heteroscedastic errors and adopt more complicated distributional structures while the link selection only considers possible different connections between the mean of the response and the covariates. An initial exploratory analysis may help to decide what is more appropriate. In other cases we may incorporate both transformation and link selection in MCMC and let the data decide which is more appropriate. The latter approach may result to selection of complicated models with no practical interpretation.

	Distribution of Y	Linear Predictor η_i	$Var(Y_i)$
No Link or Transformation	Normal	$E(Y_i)$	σ^2
Link	Normal	$g[E(Y_i)]$	σ^2
Transformation	Unknown	$g[E(Y_i)] + g''(E(Y_i))Var(Y_i)/2$	$\sigma^2/g'[E(Y_i)]$

Table 5.3: Comparison of Link and Transformation Attributes in Normal Regression.

5.4 Distribution Selection

Distribution selection can easily be handled in a similar way as link selection in the case where we consider a discrete number of distributions. Two usual cases of distribution selection problems are the normal/Student distribution and the Poisson/negative binomial.

For the selection between normal and Student's we may simply handle it as a model with Student distribution with the degrees of freedom taken as an extra parameter and then draw inferences from its posterior distribution. The likelihood of the equivalent linear model with Student distribution is given by

$$f(\boldsymbol{y}|\boldsymbol{\eta}, \sigma^2, df) = \frac{\Gamma([df+n]/2)}{\Gamma(df/2)(2\pi\sigma^2)^{n/2}} \left[1 + \frac{1}{df\sigma^2}(\boldsymbol{y}-\boldsymbol{\eta})^T(\boldsymbol{y}-\boldsymbol{\eta})\right]^{-(df+n)/2}$$

where $\boldsymbol{\eta}$ is the usual vector of linear predictors and df are the degrees of freedom. For large df the Student distribution is well approximated by the normal.

The negative binomial, can be written as a mixture Poisson distribution with

$$y_i \sim Poisson(\epsilon_i^{\delta} e^{\eta_i}), \quad \epsilon_i \sim G(\theta, \theta)$$

where δ is an indicator parameter for switching from Poisson ($\delta = 0$) to negative binomial ($\delta = 1$) setup. The parameter θ controls the over-dispersion since $E(Y_i) = e^{\eta_i}$ and $Var(Y_i) = e^{\eta_i} + e^{2\eta_i}/\theta$. Large values of θ imply low over-dispersion. The full likelihood is given by

$$f(\boldsymbol{y}, \boldsymbol{\epsilon}|\boldsymbol{\eta}, \theta, \boldsymbol{\gamma}, \delta) = f(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\epsilon}, \boldsymbol{\gamma}) f(\boldsymbol{\epsilon}|\theta, \delta).$$

For the canonical link we have

$$f(\boldsymbol{y}|\boldsymbol{\eta}, \boldsymbol{\epsilon}, \gamma) = exp\left\{-\sum_{i=1}^{n} \epsilon_i^{\delta} e^{\eta_i} + \delta \sum_{i=1}^{n} y_i log(\epsilon_i) + \sum_{i=1}^{n} y_i \eta_i - \sum_{i=1}^{n} log(y_i!)\right\}$$

and

$$f(\boldsymbol{\epsilon}|\theta, \delta = 1) exp\left\{-\theta \sum_{i=1}^{n} \epsilon_i + (\theta - 1) \sum_{i=1}^{n} log(\epsilon_i) + n\theta log(\theta) - nlog[\Gamma(\theta)]\right\}.$$

The sampling procedure is similar to simple variable selection including some parts for generating from the latent variables ϵ_i and the dispersion parameter from the corresponding posterior distributions or a suitably selected pseudoprior. The distribution selection step can be constructed using similar procedures as described in chapter for variable selection.

5.5 Illustrated Example

In this Section we illustrate various variable and link selection methods applied on the Healy (1988) dataset presented at page 131. The prior used is the same as in Section 4.6.3.1 for the logit link while the other link functions were defined using the Taylor adjustment around 0.4 which is calculated by

$$\bar{p} = \frac{1}{4}\sum_{i=1}^{2}\sum_{l=1}^{2}\left(\frac{Y_{il}}{N_{il}}\right)$$

where Y_{il} and N_{il} are the number of survivals and the total number of patients under i severity and l treatment respectively. We considered four possible link functions: the logit, complementary log-log, probit and log-log. The log-log link was used in order to illustrate convergence difficulties when one link function produces posterior parameter densities far away from the others.

Details of the methods used are provided in Table 5.4. For all methods a total of 110,000 iteration were considered discarding the first 10,000 as burn in. Reversible jump and Carlin and Chib variable and link modification performed very well and reached convergence very fast (mean batch standard deviation lower than 0.005). The simple extensions of Gibbs variable selection, SSVS and Kuo and Mallick methods did not perform well and did not reach convergence (mean batch standard deviation of 0.027, 0.053, 0.028 respectively). This is confirmed also by the barplots of Figure 5.1 where the two reversible jumps and Carlin and Chib variable and link selection (CCVLS) have very smooth distributions while the other samplers demonstrate rough changes. Note that Gibbs variable selection (GVLS) using pilot run proposal seems to have reached convergence but this is not the case. The log-log link was not visited at all in the 110,000 iterations examined due to the difficulties discussed in Section 5.1.3. On the other hand, using proposal centered at zero with pseudoprior variance 100 times smaller than the prior variance worked better but convergence was very slow compared to reversible jump or Carlin and Chib based schemes.

In the three converged methods eight out of the twenty possible models had posterior probability higher than 5%. All these models include the factor severity (A) while four of them additionally include the treatment factor (B). The model with logit link and the severity factor has the highest probability (about 14%) while the second model is given by log-log link and the severity factor with probability of of about 13%. Note that all these eight models are quite close in terms of posterior probabilities and a selection of a single model is not suggested. In such cases Bayesian model averaging is proposed instead of selecting a single model.

Figures 5.2 and 5.3 give comparisons of the ergodic posterior probabilities of the best eight models for different algorithms. In Figure 5.2 we clearly see that the reversible jumps and Carlin and Chib variable and link selection are quite close. In Figure 5.3 we compare the rest of the methods used with the first reversible jump scheme. Only GVLS with $k = 10$ seems to give close results although demonstrates greater variability. Finally, Figure 5.4 compares the ergodic probabilities when the adjustment of non-canonical link is made by Taylor expansion around $\bar{p} = 0.4$, Taylor expansion around 0.5 and maximum likelihood standard deviations. Taylor approximation around 0.4 and Maximum likelihood estimates adjustment gave similar results in clog-log and log-log link functions while two the Taylor based approximations gave similar results for in the probit link. This may indicates that for the probit link the Taylor expansion round 0.5 is sufficient (probably due to the symmetry of the link) while for the other two (asymmetric) links more complicated adjustments are demanded (based on maximum likelihood estimates or Taylor expansion round $\bar{p}$).

		Algorithm for Selection of			
	Abbreviation	Variable		Link	
1	RJ_1	Simple RJ	(eq. 4.8-4.9)	RJ with Transformation	(eq. 5.9)
		[Transformation: eq. 5.7 & 5.8 with $p = \bar{p}$]			
2	RJ_2	Simple RJ	(eq. 4.8-4.9)	RJ with Pilot Run Proposals	(eq. 5.10)
3	CCVLS	GVS	(eq. 4.3)	Carlin and Chib Type	(eq. 5.12)
4	GVLS ($k = 10$)	GVS	(eq. 4.3)	Simple Gibbs Step	(eq. 5.11)
		[Pseudoprior: $N(0, \Sigma_j/k^2)$ for Each Link]			
5	SSVLS	SSVS	(eq. 3.25)	Simple Gibbs Step	(eq. 5.11)
6	KMVLS	KM	(eq. 3.30)	Simple Gibbs Step	(eq. 5.11)
7	GVLS	GVS	(eq. 3.30)	Simple Gibbs Step	(eq. 5.11)
		[Pseudoprior: Full Model Pilot Run for Each Link]			

Table 5.4: Variable and Link Algorithms Used in Healy Data.

Model		Posterior Probability			
Factors	Link	RJ_1	RJ_2	CCVLS	GVLS ($k=10$)
A	logit	0.139	0.138	0.137	0.140
A	log-log	0.130	0.131	0.130	0.134
A	probit	0.127	0.127	0.126	0.126
A+B	logit	0.125	0.125	0.130	0.123
A	cloglog	0.121	0.121	0.122	0.121
A+B	cloglog	0.121	0.121	0.122	0.121
A+B	probit	0.100	0.101	0.103	0.096
A+B	log-log	0.073	0.072	0.077	0.078

Table 5.5: Link and Variable Selection for Healy Data: Posterior Model Probabilities Higher than 0.05; See Table 5.4 for Acronyms.

	Posterior Probability			
Link	RJ_1	RJ_2	CCVLS	GVLS ($k=10$)
Logit	0.284	0.283	0.282	0.282
Cloglog	0.253	0.254	0.257	0.251
Probit	0.243	0.244	0.242	0.238
Loglog	0.220	0.220	0.219	0.230

Table 5.6: Link and Variable Selection in Healy Data: Marginal Posterior Distribution of Link Functions; See Table 5.4 for Acronyms.

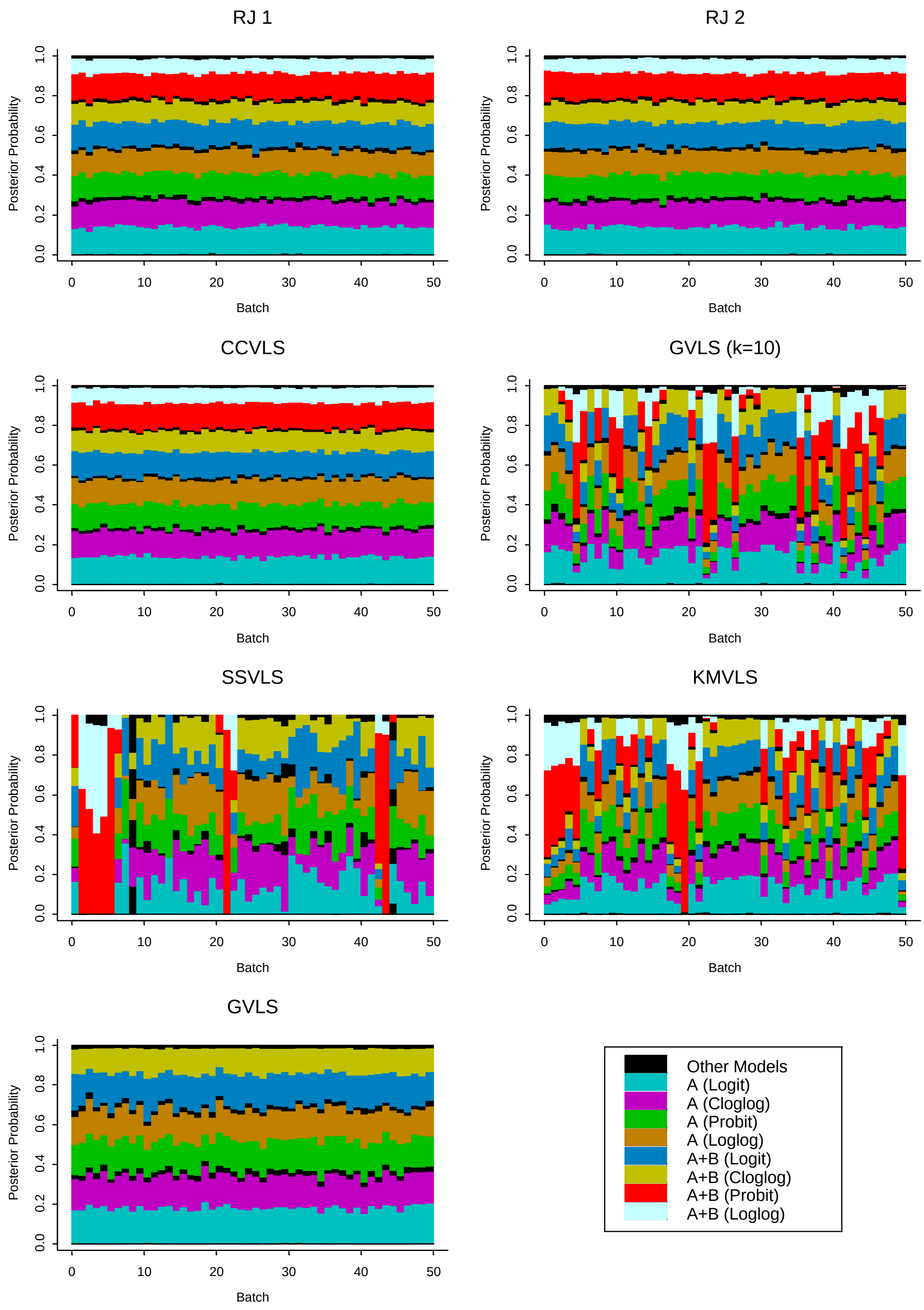


Figure 5.1: Link and Variable Selection in Healy Data: Batch Posterior Model Probabilities of Different Variable and Link Selection Methods; See Table 5.4 for Acronyms.

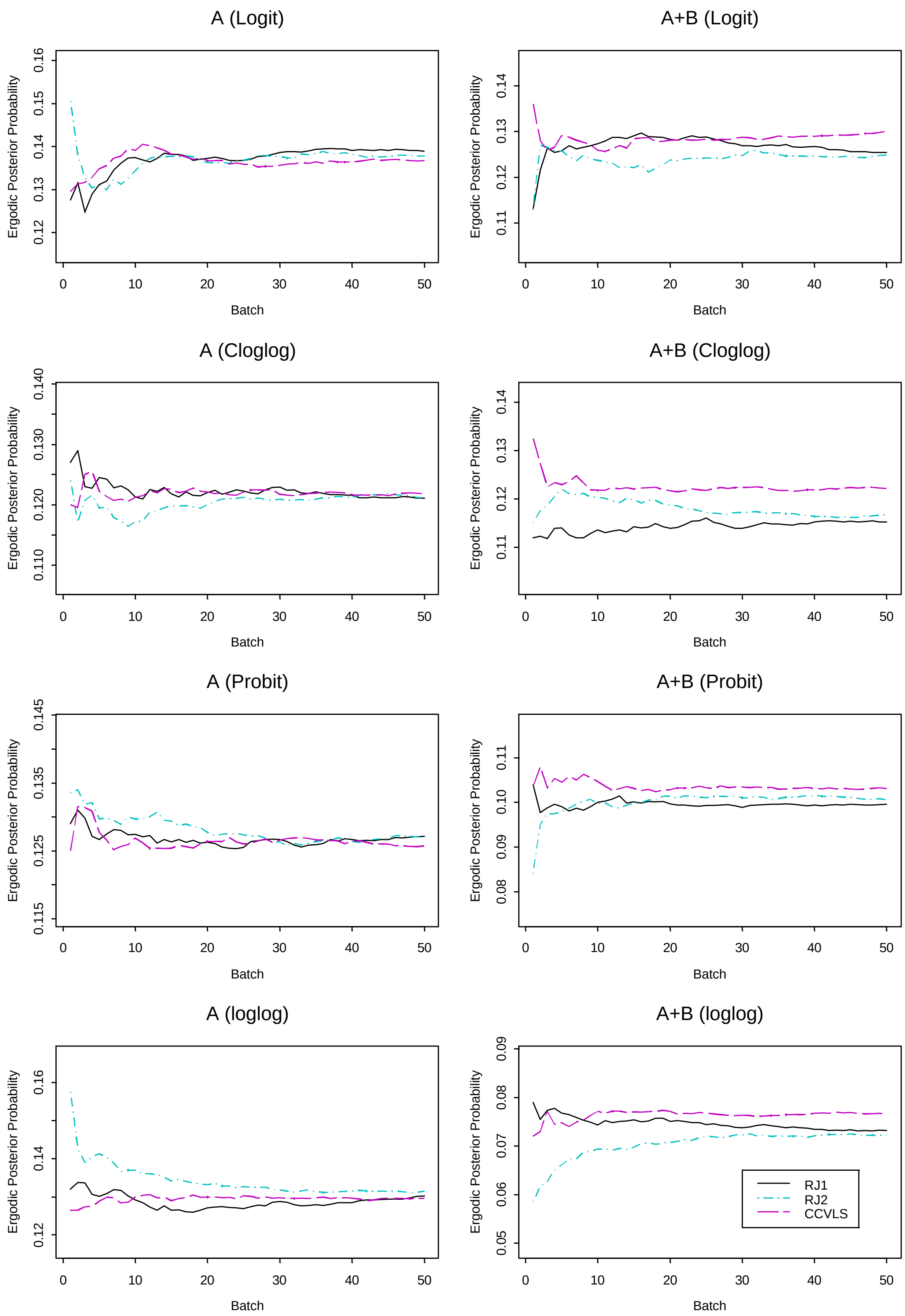


Figure 5.2: Link and Variable Selection in Healy Data: Ergodic Posterior Model Probabilities of RJ_1, RJ_2 and CCVLS; See Table 5.4 for Acronyms.

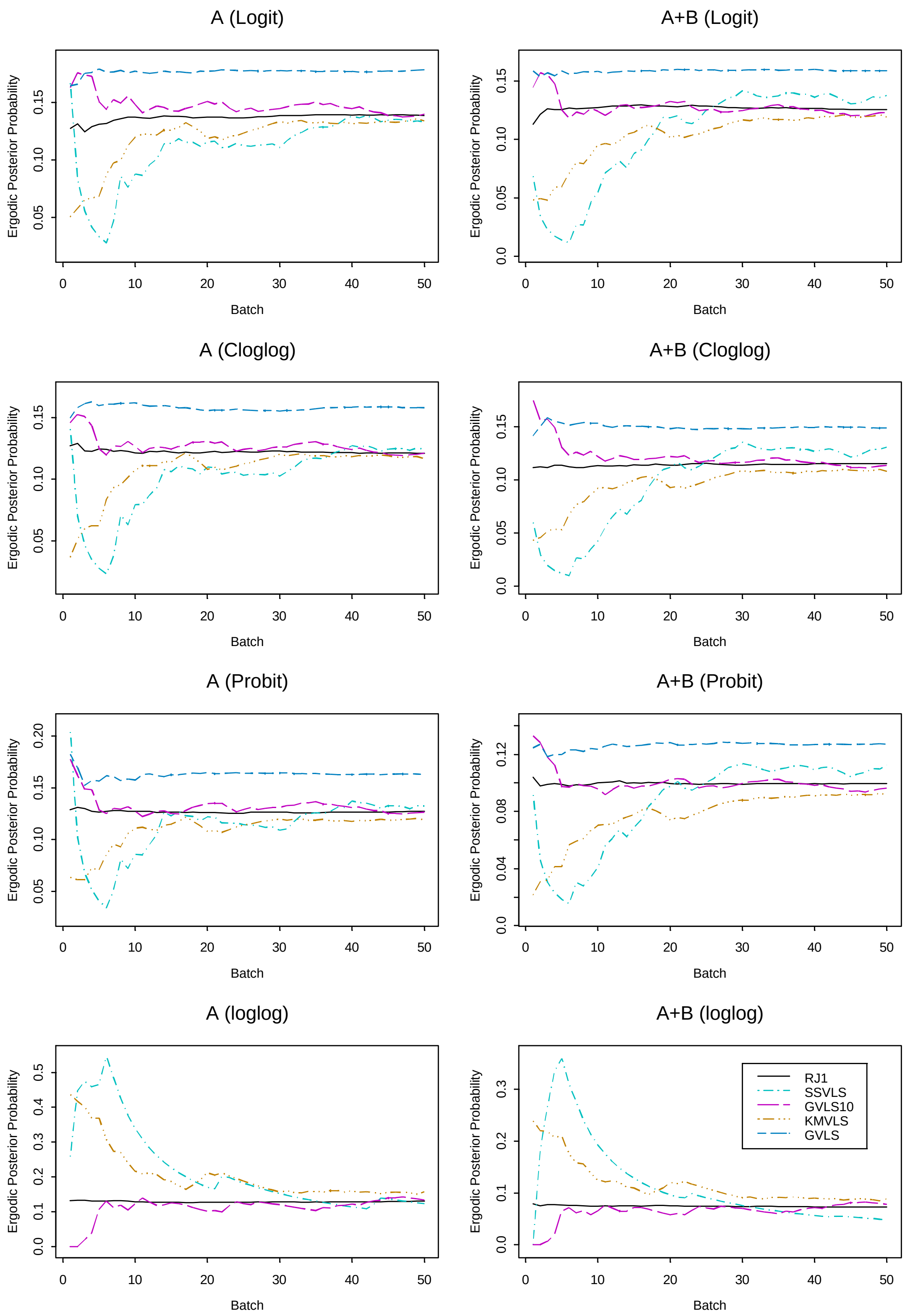


Figure 5.3: Link and Variable Selection in Healy Data: Ergodic Posterior Model Probabilities Comparison of RJ_1, SSVLS, GVLS (k=10), KMVLs and GVLS; See Table 5.4 for Acronyms.

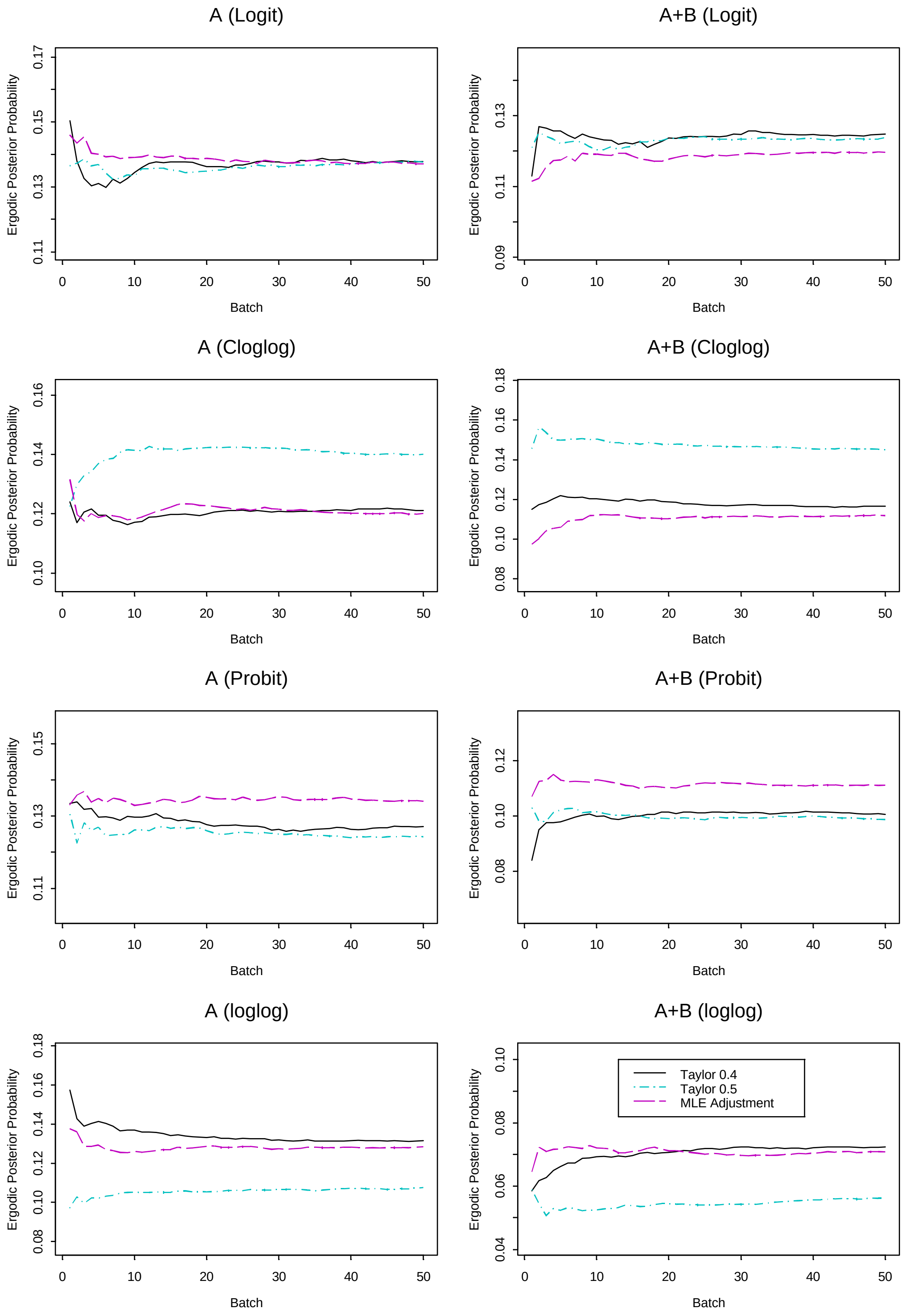


Figure 5.4: Link and Variable Selection in Healy Data: Comparison of Ergodic Posterior Model Probabilities for Different Prior Distribution Link Adjustment.

Chapter 6

On Prior Distributions for Model Selection

6.1 Introduction

In this chapter we consider the problem of prior specification for model selection with emphasis on the normal linear model. The main problem in prior specification when no information is available is that we cannot use improper priors on model parameters due to the unknown normalizing constants involved in the calculation of the posterior odds. Moreover, even if we use proper prior distributions on model parameters, the posterior odds are very sensitive on the magnitude of the prior variance, tending to support the simpler model as the prior variance increases (Lindley-Bartlett paradox, 1957). We would like to find a prior distribution that will be non-informative *within each model* (in the sense that the posterior modes will be close to the maximum likelihood estimates), and coherent in terms of dimension penalty imposed *between models*.

In Section 6.2 we briefly describe the general model comparison in linear model and two simple examples that motivated our work. We also propose that $-2log(PO_{01})$, where PO_{01} is the posterior odds of model m_0 against model m_1, can be expressed in a similar way as information criteria which can be divided into two parts. The first one can be called as ratio of posterior sum of squares (in analogy to the maximum likelihood residual sum of squares used in most information criteria) and a penalty function which depends on the prior variance of model parameters and the prior model probabilities. We further investigate whether we can use independent prior distributions in variable selection setups and we argue that in collinear cases this leads to paradoxes. In Section 6.3 we argue that the usual uniform distribution on the model space leads to incoherent prior distributions if we use alternative measures such as conditional prior odds at zero which are based on an idea of Robert (1993). In Section 6.4 we propose a prior specification technique which enables us to eliminate the prior variance effect on the posterior odds, to impose the penalty we prefer for each additional parameter included in the model, and to achieve the desired coherency. Following Lindley's (1957) arguments we investigate the behaviour of posterior odds in the limit of significance in normal linear model in Section 6.5. In Section 6.6 we present how we can specify prior distributions with low information via penalty determination in generalised linear models. Finally, in Section 6.7 we discuss the behaviour Bayes factor variants including the SSVS based Bayes factor. Note that some early results of this chapter have been presented in Ntzoufras *et al.* (1999a).

6.2 The Normal Linear Model

6.2.1 A General Model Comparison

Let us consider two models m_0 and m_1 given by

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n\right),$$

where $m \in \{m_0, m_1\}$, n is the sample size, $\boldsymbol{\beta}_{(m)}$ is the $d(m) \times 1$ vector of model parameters, $\boldsymbol{X}_{(m)}$ is the $n \times d(m)$ design (or data) matrix of model m, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}$ and $\boldsymbol{I}_n$ is $n \times n$ identity matrix.

We adopt the conjugate prior distribution given by

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, c^2 \boldsymbol{V}_{(m)}\sigma^2\right)$$

and the improper prior for the residual variance $f(\sigma^2) \propto \sigma^{-2}$.

The resulting posterior odds for comparing model m_0 with model m_1 are given by

$$PO_{01} = c^{d(m_1)-d(m_0)} \left(\frac{|\boldsymbol{V}_{(m_0)}|}{|\boldsymbol{V}_{(m_1)}|}\right)^{-1/2} \left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|}{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|}\right)^{1/2} \left(\frac{SS_{m_0}}{SS_{m_1}}\right)^{-n/2} \frac{f(m_0)}{f(m_1)} \qquad (6.1)$$

where SS_m are the posterior sum of squares given by

$$SS_m = \boldsymbol{y}^T\boldsymbol{y} - \tilde{\boldsymbol{\beta}}^T_{(m)}\tilde{\boldsymbol{\Sigma}}^{-1}_{(m)}\tilde{\boldsymbol{\beta}}_{(m)} + c^{-2}\boldsymbol{\mu}^T_{\beta_{(m)}}\boldsymbol{V}^{-1}_{(m)}\boldsymbol{\mu}_{\beta_{(m)}}, \tag{6.2}$$

$$\tilde{\boldsymbol{\beta}}_{(m)} = \tilde{\boldsymbol{\Sigma}}_{(m)}\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)} + c^{-2}\boldsymbol{V}^{-1}_{(m)}\boldsymbol{\mu}_{\beta_{(m)}}\right),\ \tilde{\boldsymbol{\Sigma}}^{-1}_{(m)} = \boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{V}^{-1}_{(m)}, \tag{6.3}$$

$\boldsymbol{\mu}_{\beta_{(m)}}$ and $\boldsymbol{V}_{(m)}$ are the prior mean vector and a prior matrix associated with the prior variance-covariance matrix of the parameter vector $\boldsymbol{\beta}_{(m)}$ respectively, and $\tilde{\boldsymbol{\beta}}_{(m)}$ and $\tilde{\boldsymbol{\Sigma}}_{(m)}$ are the corresponding posterior measures.

An alternative expression of the posterior sum of squares is given by Atkinson (1978) and Pericchi (1984) where

$$SS_m = RSS_m + \left(\hat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\mu}_{\beta_{(m)}}\right)^T\left[\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\right)^{-1} + c^2\boldsymbol{V}_{(m)}\right]^{-1}\left(\hat{\boldsymbol{\beta}}_{(m)} - \boldsymbol{\mu}_{\beta_{(m)}}\right) \tag{6.4}$$

with $RSS_m = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T_{(m)}\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\hat{\boldsymbol{\beta}}_{(m)}$ being the usual residual sum of squares. The above quantity is the sum of two measures: a goodness of fit and a measure of distance between maximum likelihood estimates and the prior mean.

6.2.2 Motivation

Our motivation for searching a suitable prior emerges from our need to use:

1. Non-informative priors within each model.

2. Independent priors on model parameters.

Unfortunately, when we try to use either non-informative or independent prior distributions in model selection, paradoxes emerge. In the first case we have the Lindley-Bartlett paradox while in the second case the posterior probabilities of two collinear models are close; see Section 6.2.4 at page 166.

Example One

Consider the model $\boldsymbol{y} \sim N(\beta_1\boldsymbol{X}_1, \boldsymbol{I}_n)$ with a normal prior distribution on β_1 given by $N(\mu_{\beta_1}, \sigma^2_{\beta_1})$. Then the posterior density is given by

$$f(\beta_1|\boldsymbol{y}) \sim N\left(w\hat{\beta}_1 + (1-w)\mu_{\beta_1}, wVar(\hat{\beta}_1)\right),\quad Var(\hat{\beta}_1) = \left(\sum_{i=1}^n X^2_{i,1}\right)^{-1},$$

where $w = \sigma^2_{\beta_1}\left(\sigma^2_{\beta_1} + Var(\hat{\beta}_1)\right)^{-1}$. Consider now a simpler normal prior with $\mu_{\beta_1} = 0$ and $\sigma^2_{\beta_1} = c^2[\sum_{i=1}^n X^2_{i,1}]^{-1}$. The posterior is simplified to

$$f(\beta_1|\boldsymbol{y}) \sim N\left(\frac{c^2}{c^2+1}\hat{\beta}_1, \frac{c^2}{c^2+1}Var(\hat{\beta}_1)\right).$$

In such case it is natural to assume that $c^2 = 1000$ can be considered as non-informative since $w = 0.999 \approx 1$. This is not always the case. Consider a case where the z statistic is very large, for example $\hat{\beta}_1 = 10^6$ and $Var(\hat{\beta}_1) = 10$ [$\sum_{i=1}^n X^2_{i,1} = 1/10$ and $\sum_{i=1}^n Y_i X_{i,1} = 10^5$]. Then a choice of $c^2 = 1000$ will result to a posterior density

$$f(\beta_1|\boldsymbol{y}) \sim N\left(999001, 10\right),$$

while for improper flat prior the posterior density is equal to

$$f(\beta_1|\boldsymbol{y}) \sim N\left(1000000, 10\right).$$

The two densities are far away and therefore the choice of $c^2 = 1000$ cannot be thought as non-informative. Clearly, if such a prior is used for one model parameter the results from a Bayesian model averaging approach will contain prior information that influences the posterior densities.

Example Two: Collinearity, Independent Priors and Model Selection

Consider two normal models $m_0 : \boldsymbol{y} \sim N(\beta_1 \boldsymbol{X}_1, \boldsymbol{I}_n\sigma^2)$ and $m_1 : \boldsymbol{y} \sim N(\beta_1\boldsymbol{X}_1 + \beta_2\boldsymbol{X}_2, \boldsymbol{I}_n\sigma^2)$ with priors

$$\beta_j \sim N\left(0, c^2\left[\sum_{i=1}^n X^2_{ij}\right]^{-1}\sigma^2\right) \text{ for } j = 1, 2.$$

In the extreme (collinear) case where $X_2 \propto X_1$, for large c^2 and equal prior model probabilities, we have that the posterior odds are equal to 1.41 in favour of the simplest model whatever the relationship of X_1 and Y is! More details of this motivated example are given in Section 6.2.4 (page 166). This paradox is true for all Y and X and clearly indicates that independent prior distributions may be inappropriate for regression models.

6.2.3 Posterior Odds and Information Criteria

Generally, most information criteria select the model that minimizes a quantity which is usually equal to the maximum likelihood ratio penalized for each additional term used in the model; see equation (2.14). There are many model selection criteria usually characterized by the type of penalties; detailed discussion is presented in Section 2.3. In pairwise comparisons of models m_0 and m_1 we will use the difference of two information criteria $IC_{01} = IC_0 - IC_1$ defined by equation (2.14).

It is obvious that if $IC_{01} < 0$ we prefer the simpler model m_0 and when $IC_{01} > 0$ we prefer the more complicated model m_1. Further note that we have the same support pattern when we use the minus twice the logarithm of the posterior odds of model m_0 vs. model m_1 denoted by PO_{01} since $-2log(PO_{01}) < 0$ supports the simpler model m_0 while $-2log(PO_{01}) > 0$ supports the more complicated model m_1. Moreover, there is a variety of publications connecting specific information criteria with posterior odds and Bayes factor; see Schwarz (1978), Smith and Spiegelhalter (1980), Poskitt and Tremayne (1983) and Kass and Wasserman (1995).

Here we remind that we may write the information criteria in a general setup given by

$$IC_{01} = -2log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_0}, m_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_1}, m_1)}\right) - \psi, \tag{6.5}$$

where ψ is a penalty function depending on the difference of model dimensionalities $d(m_1) - d(m_0)$, the sample size n, and the design matrices $\boldsymbol{X}_{(m_0)}$ and $\boldsymbol{X}_{(m_1)}$. Also in normal linear models the above quantity is simplified to

$$IC_{01} = nlog\left(\frac{RSS_{m_0}}{RSS_{m_1}}\right) - \psi, \tag{6.6}$$

where RSS_m are the residual sum of squares of model m.

In the model comparison proposed in Section 6.2.1, $-2log(PO_{01})$ can be written as

$$-2log(PO_{01}) = nlog\left(\frac{SS_{m_0}}{SS_{m_1}}\right) - \psi, \tag{6.7}$$

$$\psi = \{\, d(m_1) - d(m_0)\,\}log(c^2) + log\left(\frac{|\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{V}_{(m_0)}|}\right) + log\left(\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}^{-1}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}^{-1}_{(m_0)}|}\right) - 2log\left(\frac{f(m_1)}{f(m_0)}\right). \tag{6.8}$$

Note that if we consider equation (6.4) then it is obvious that the posterior sum of squares incorporate both prior and data information. In addition, if we consider large c^2 then posterior sum of squares becomes approximately equal to the usual residual sum of squares.

Proposition 6.1 *Consider two normal linear models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n\right)$$

and prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2\boldsymbol{V}_{(m)}\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$. Then $-2log(PO_{01})$, for large c^2, is approximately equal to an information criterion of type (6.5) with penalty function

$$\psi = \{\ d(m_1) - d(m_0)\ \}log(c^2) + log\left(\frac{|\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{V}_{(m_0)}|}\right) + log\left(\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}|}\right) - 2log\left(\frac{f(m_1)}{f(m_0)}\right). \quad (6.9)$$

Proof: See appendix, page 199. $\triangleleft$

From the above proposition, it is evident that, for large c^2, the $-2log(PO_{01})$ is equivalent to an information criterion with penalty affected by the prior parameter c^2. What is desirable is to eliminate the effect of c^2 from the penalty imposed to the ratio of posterior odds. Furthermore, adopting the prior distribution suggested by Smith and Kohn (1996) leads to the following proposition.

Proposition 6.2 *Consider two normal linear models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n\right)$$

and prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\right)^{-1}\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2} \quad (6.10)$$

for both $m \in \{m_0, m_1\}$. Then $-2log(PO_{01})$ is equal to an information like criterion given by (6.7) and, for large c^2, approximately equal to an information criterion of type (6.5) with penalty function

$$\psi = \{d(m_1) - d(m_0)\}log(c^2 + 1) - 2log\left(\frac{f(m_1)}{f(m_0)}\right). \quad (6.11)$$

Proof: See appendix, page 199. $\triangleleft$

The above prior setup results in a Bayes factor that allows us to fully control the penalty of the log-likelihood via c^2. Both c^2 and the prior odds have different influence on the penalty (and the posterior odds). The above result was independently reported by Fernandez *et al.* (1998).

This prior cannot be computed if an explanatory variable is an exact linear function of the others. In such extreme case we should remove this term from the model. Consider now the case of the comparison between two models that differ only by one explanatory variable involving a single parameter. When the additional variable is highly correlated with the other variables in the model m_0 (e.g. regression $R^2 = 0.99$ of the new variable over the others) then the likelihood ratio test will be very small and therefore the Bayes factor will be close to $c^2 + 1$ in favour of the simpler model. Therefore, the prior parameter c^2 in this prior setup has two straightforward interpretations. The first is closely related with the penalty attributed to the log-likelihood ratio for each additional parameter used and the second is related with the Bayes factor of the model m_0 vs. the same model with one additional variable which is highly correlated with the rest variables in the model.

6.2.4 Independent Prior Distributions for Variable Selection

In linear models it is convenient to use the approach of variable selection. For this reason we can express the data matrix of the full model $\boldsymbol{X}$ and its parameter vector $\boldsymbol{\beta}$ as a collection of distinct $n \times d_j$ sub-matrices $\boldsymbol{X}_j$ and $d_j \times 1$ sub-vectors $\boldsymbol{\beta}_j$, each one corresponding to j term (factor, variable or interaction term); where d_j are the number of parameters involved in j term. Therefore we can write

$$\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p] \text{ and } \boldsymbol{\beta}^T = [\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_p^T],$$

where p is the number of terms in the full model. Under such consideration, a convenient setup is to use independent normal priors on these sub-vectors. Moreover, each design matrix $\boldsymbol{X}_{(m)}$ and each parameter vector $\boldsymbol{\beta}_{(m)}$ is constructed by the corresponding sub-matrices $\boldsymbol{X}_j$ and sub-vectors $\boldsymbol{\beta}_j$ for all terms j included in model m.

Suppose we use the priors

$$f(\boldsymbol{\beta}_j|\sigma^2) \sim N\left(0, c^2(\boldsymbol{X}_j^T\boldsymbol{X}_j)^{-1}\sigma^2\right)$$

for all $j \in \mathcal{V}(m)$, where $\mathcal{V}(m)$ is the set of all terms included in model m, $m \in \mathcal{V}$. Raftery *et al.* (1997) proposed similar priors for generalised linear models. To facilitate and link the connection between model and variable selection we use a vector of variable indicators $\boldsymbol{\gamma}_m$ denoting which terms are included in model m and $\gamma_{j,m}$ is the value of the indicator of j

term for model m. The prior of each $\boldsymbol{\beta}_{(m)}$ is given by the product of all priors $f(\boldsymbol{\beta}_j|\sigma^2)$ for the terms included in model m (for which $\gamma_{j,m} = 1$).

Let us firstly consider a simple model comparison of two models m_0 and m_1 in which m_1 differs from m_0 by only one additional term j. Then we have the following proposition.

Proposition 6.3 *Consider two nested normal models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ and additionally that

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j]$$

(that is $\gamma_{j',m_0} = \gamma_{j',m_1}$ for all $j' \neq j$ and $\gamma_{j,m_0} = 0$, $\gamma_{j,m_1} = 1$) with independent normal prior distributions that can be summarised by

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2} \tag{6.12}$$

where $D_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)$ is a $d(m) \times d(m)$ block diagonal matrix with elements the matrices $\boldsymbol{X}^T_j\boldsymbol{X}_j$ for all j terms included in model m. Then the resulted penalty is given by

$$\psi = d_j log(c^2+1) - log|\boldsymbol{X}^T_j\boldsymbol{X}_j| + log|\boldsymbol{X}^T_j\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j| + 2log\left(\frac{f(m_0)}{f(m_1)}\right), \tag{6.13}$$

$$\boldsymbol{\Delta}_{(m_0)} = \boldsymbol{I}_n - \frac{c^2}{c^2+1}\boldsymbol{X}_{(m_0)}\tilde{\boldsymbol{\Sigma}}_{(m_0)}\boldsymbol{X}^T_{(m_0)}, \ \ \tilde{\boldsymbol{\Sigma}}^{-1}_{(m_0)} = \boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}^T_j\boldsymbol{X}_j).$$

Proof: See appendix, page 200. $\triangleleft$

Furthermore, if the design matrix of the full model is orthogonal we have the following corollary.

Corollary 6.3.1 *Consider two nested normal models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ with independent normal prior distributions that can be summarised by

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

where $D_{(m)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)$ is a $d(m)\times d(m)$ block diagonal matrix with elements the matrices $\boldsymbol{X}_j^T\boldsymbol{X}_j$ for all j terms included in model m. If the design (or data) matrix of the additional j term is orthogonal to the design (or data) matrix of the null model then the penalty degenerates to

$$\psi = \{\ d(m_1) - d(m_0)\ \}log(c^2+1) + 2log\left(\frac{f(m_0)}{f(m_1)}\right). \qquad (6.14)$$

Proof: See appendix, page 201. $\triangleleft$

Further, consider the case that X_j is a vector corresponding to one dimensional term then the regression model

$$\boldsymbol{X}_j \sim N(\boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0,x_j)}, \sigma^2\boldsymbol{I}_n)$$

using independent normal prior distributions for the model parameters for $\boldsymbol{\beta}_{(m_0,x_j)}$ similar to (6.12) given by

$$f(\boldsymbol{\beta}_{(m_0,x_j)}|\sigma^2, m_0) \sim N\left(\boldsymbol{0}, c^2 D^{-1}_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}.$$

For this new model, the posterior sum of squares is given by

$$SS_{m_0,x_j} = \boldsymbol{X}_j^T\boldsymbol{X}_j - \boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\left[\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right]^{-1}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j.$$

Therefore for large c^2

$$\begin{aligned}
\boldsymbol{X}_j^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j &= \boldsymbol{X}_j^T\boldsymbol{X}_j - \frac{c^2}{c^2+1}\boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\left[\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right]^{-1}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j \\
&\approx \boldsymbol{X}_j^T\boldsymbol{X}_j - \boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\left[\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right]^{-1}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j \\
&\approx SS_{m_0,x_j}.
\end{aligned}$$

It is obvious that the determinant $|\boldsymbol{X}_j^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j|$ is a measure of collinearity between the additional term j and the terms of model m_0. When X_j is orthogonal to $X_{(m_0)}$ then the above determinant and the corresponding penalty are maximized resulting to (6.14). On the other hand, the corresponding penalty of proposition 6.3 is minimized when X_j is collinear to the terms of model m_0 since the above determinant will be close to zero. Furthermore, in the collinear case the ratio of the sum of squares will be equal to one and therefore the Bayes factor will depend only on the imposed penalty ($\psi = d_j log(c^2+1) - log|\boldsymbol{X}_j^T\boldsymbol{X}_j| + 2log\left(\frac{f(m_0)}{f(m_1)}\right)$). What is desirable and plausible is that when X_1 and X_2 are collinear and $X_{(m)}$ denotes a collection of terms that define model m for any $m \in \mathcal{M}$, then a Bayes factor of any model with one of the collinear variables $m + X_1$ (or $m + X_2$) versus the same model with the two collinear variables, $m + X_1 + X_2$, should fully support the simpler model. On the other hand, any comparison of the $m + X_1$ and $m + X_2$ should result to a Bayes factor equal to one since both models have the same posterior distributions. This is true since the dimensionality difference of the two models is equal to zero and the posterior sum of squares are the same resulting in Bayes factor equal to one. For this reason, we should be careful and try to satisfy the first plausible statement (fully support models $m + X_1$ against models $m + X_1 + X_2$ for all $m \in \mathcal{M}$). The example which follows clearly demonstrates that this statement is not satisfied when independent prior distributions are used.

A Simple Example

In this example we use the simplest possible case where $d(m_0) = 1$ and $d(m_1) = 2$ in order to illustrate the effect of such priors in model selection. We compare a model m_0 with $\boldsymbol{X}_{(m_0)} = [\boldsymbol{X}_1]$ with parameter vector $\boldsymbol{\beta}_{(m_0)} = [\beta_1]$ and a model m_1 with $\boldsymbol{X}^T_{(m_1)} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{\beta}_{(m_1)} = [\beta_1, \beta_2]$. The resulting penalty (6.13) using the $f(m_0) = f(m_1)$ is equal to

$$\psi = log(c^2 + 1) + log\left(1 - \left(\frac{c^2}{c^2+1}\right)^2 r^*_{x_1x_2}\right)$$

with

$$r^*_{x_1x_2} = \frac{\left(r_{x_1x_2} + \frac{\bar{x}_1}{\sigma_{x_1}}\frac{\bar{x}_2}{\sigma_{x_2}}\right)^2}{\left(1 + \frac{\bar{x}_1^2}{\sigma^2_{x_1}}\right)\left(1 + \frac{\bar{x}_2^2}{\sigma^2_{x_2}}\right)} = \frac{\left(r_{x_1x_2} + \frac{\bar{x}_1}{\sigma_{x_1}}\frac{\bar{x}_2}{\sigma_{x_2}}\right)^2}{\left(1 + \frac{\bar{x}_1}{\sigma_{x_1}}\frac{\bar{x}_2}{\sigma_{x_2}}\right)^2 + \left(\frac{\bar{x}_1}{\sigma_{x_1}} - \frac{\bar{x}_2}{\sigma_{x_2}}\right)^2} \leq 1$$

where $r_{x_1x_2}$ are correlation coefficients of X_1 and X_2 with $r_{x_1x_2} \leq 1$ and $\sigma^2_{x_j}$ are the biased estimates of variance for X_j variable .

Consider the extreme case where $X_2 \propto X_1$; then $r_{x_1x_2} = 1$ and $r^*_{x_1x_2} = 1$. In such a case both variables carry exactly the same information and only one of them should be included in the model. The above setup results to penalty $\psi = log\left(\frac{2c^2+1}{c^2+1}\right)$. For large c, $\psi \approx log(2)$ resulting in Bayes factor equal to 1.41 (58.5% over 41.5%) in favour of the simpler model whatever is the relation between X_1 and Y! Similar is the case in which the parameter vector of each term is multidimensional. Note that Aitkin's posterior odds (1991) give penalty equal to $log(2)$ to the log-likelihood; see Aitkin (1991) and O'Hagan (1995).

Possible ways to avoid the effect of collinearity on posterior odds are to use the orthogonal transformed data, see Clyde *et al.* (1996), Clyde (1999), or use the model dependent priors used by Smith and Kohn (1996) and George and Foster (1997); see Section 6.2.1 for details.

We may generalise the result of the above example for any one dimensional additional term X_j. In these cases all matrices involved in the penalty function (6.13) are of 1×1 dimension and therefore we can omit all determinants.

Proposition 6.4 *Consider two nested normal models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ and additionally that

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j]$$

(that is $\gamma_{j',m_0} = \gamma_{j',m_1}$ for all $j' \neq j$ and $\gamma_{j,m_0} = 0,\ \gamma_{j,m_1} = 1$) with independent normal prior distributions that can be summarised by the prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

and X_j one dimensional term, collinear to the terms included in model m_0 then, for large c^2, $2log(PO_{01})$ is approximately equal

$$2log(PO_{01}) \approx \psi = log[R^*(m_0, X_j) + 1] + 2log\left(\frac{f(m_0)}{f(m_1)}\right)$$

where

$$R^*(m_0, x_j) = \frac{\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)} D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\hat{\boldsymbol{\beta}}_{(m_0,x_j)}}{\boldsymbol{X}_j^T\boldsymbol{X}_j}, \tag{6.15}$$

$$= \frac{\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)} D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j) clear \hat{\boldsymbol{\beta}}_{(m_0,x_j)}}{\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0,x_j)}} \tag{6.16}$$

$$= 1 - \frac{\alpha_1^*}{\alpha_1^* + \alpha_2^*} \tag{6.17}$$

with

$$\alpha_1^* = \sum_{\nu \in \mathcal{V}(m_0)} \sum_{\nu' \in \mathcal{V}(m_0)\setminus\{\nu\}} [\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]^T_\nu \boldsymbol{X}^T_\nu \boldsymbol{X}_{\nu'} [\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_{\nu'}$$

$$\alpha_2^* = \sum_{\nu \in \mathcal{V}(m_0)} [\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]^T_\nu \boldsymbol{X}^T_\nu \boldsymbol{X}_\nu [\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_\nu,$$

where $[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_\nu$ *are the maximum likelihood estimates for term* ν *of a normal linear model with* X_j *as response and* $\boldsymbol{X}_{(m_0)}$ *design matrix and* $\mathcal{V}(m)$ *is the set of terms included in model* m.

Proof: See appendix, page 201. $\triangleleft$

In the above proposition we clearly see that when we use independent prior distributions on model terms and compare two nested models that differ only by j term then the posterior odds depend only on the quantity $R^*(m_0, x_j)$ or the ratio $\alpha_1^*/(\alpha_1^* + \alpha_2^*)$ whatever the relationship between X_j and Y. Additionally, $R^*(m_0, x_j)$ takes values between zero and one and therefore the Bayes factor will be between 1 and 1.41 (the posterior probability, when equal weights are assumed, is between 0.500 and 0.586) in favour of the simpler model. This is clearly a paradox since, in such case, any model selection procedure should fully support the simpler model which carries the same information as the most complicated one. This is a serious drawback of the independent priors setup. The following corollary considers the case where the maximum of the penalty is observed under the extreme case of collinearity and it is a more general case of the simple example presented above.

Corollary 6.4.1 *Consider the model comparison of proposition 6.4 and additionally assume that the design matrix of the null model is orthogonal then, for large* c^2,

$$2log(PO_{01}) \approx \psi = log(2) + 2log\left(\frac{f(m_0)}{f(m_1)}\right). \tag{6.18}$$

Proof: See appendix, page 204. $\triangleleft$

6.3 Conditional Prior Odds at Zero

In Bayesian model selection it is a usual practice to adopt the uniform prior on model space as a 'non-informative' prior distribution and then specify the prior distribution of each parameter vector conditionally on the model. We argue that this uniform prior on model space leads to incoherency if we use measures of conditional prior odds based on the ideas proposed by Robert (1993).

Robert (1993) extensively discussed Lindley's paradox and argued that increasing the prior variance of a parameter is essentially giving more prior weight to the simpler hypothesis.

He examined the simple case with $Y_i \sim N(\mu, \sigma^2)$ for $i = 1, \dots n$ where

$$H_0 : \mu = 0 \;\; vs. \;\; H_1 : \mu \neq 0.$$

He used the term 'conditional prior odds at a neighbourhood of zero' $[-\xi, \xi]$ $(CPONZ_\xi)$ which is defined as following.

Definition 6.1 *The 'conditional prior odds at a neighbourhood of zero'* $[-\xi, \xi]$ $(CPONZ_\xi)$ *is given by the ratio*

$$CPONZ_\xi = \frac{f(m_1 \mid -\xi < \mu < \xi)}{f(m_0 \mid -\xi < \mu < \xi)}.$$

We adopt a similar approach for normal linear and generalised linear models. The case of model selection in these models is far more complicated than the simple case illustrated by Robert (1993).

Consider two nested normal models m_0 and m_1 given by

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ and additionally that

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j].$$

The hypothesis test may be written as

$$H_0 : \boldsymbol{\beta}_j = \mathbf{0} \;\; vs. \;\; H_1 : \boldsymbol{\beta}_j \neq \mathbf{0}.$$

We can also facilitate the use of a variable indicator γ_j and summarize both models by $\boldsymbol{\mu}_{(m_{\gamma_j})} = \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$ for $\gamma_j = 0, 1$. Instead of the usual prior odds $f(m_1)/f(m_0) = f(\gamma_j = 1)/f(\gamma_j = 0)$ we may adopt the approach of 'conditional prior odds at a neighbourhood of zero' $[-\xi, \xi]$ which is given by the following definition.

Definition 6.2 *The conditional prior odds for comparing two normal linear models* $\boldsymbol{y} \sim N(\boldsymbol{\mu}_{(m_{\gamma_j})}, \sigma^2 \boldsymbol{I}_n)$ *with* $m_{\gamma_j} : \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma_j \boldsymbol{X}_j \boldsymbol{\beta}_j$ *for* $\gamma_j \in \{0, 1\}$ *is defined as*

$$CPONZ_\xi = \frac{f(m_1 \mid -\xi < \beta_j < \xi, \sigma^2)}{f(m_0 \mid -\xi < \beta_j < \xi, \sigma^2)} \tag{6.19}$$

when j term is one dimensional and by

$$CPONZ_\xi = \frac{f(m_1 \mid \boldsymbol{\beta}_j^T \boldsymbol{\beta}_j < \xi^2, \sigma^2)}{f(m_0 \mid \boldsymbol{\beta}_j^T \boldsymbol{\beta}_j < \xi^2, \sigma^2)} \tag{6.20}$$

when j term is multi-dimensional.

The above definition can be generalised by using a more general condition given by $\mathcal{F}(\boldsymbol{\beta}_j) < \xi'$ instead of the neighbourhood defined by the condition $\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2$; where $\mathcal{F}$ is a function defining the neighbourhood of zero.

Proposition 6.5 *The conditional prior odds of definition 6.2 for comparing two normal linear models $\boldsymbol{y} \sim N(\boldsymbol{\mu}_{(m_{\gamma_j})}, \sigma^2\boldsymbol{I}_n)$ with $\boldsymbol{\mu}_{(m_{\gamma_j})} = \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma_j\boldsymbol{X}_j\boldsymbol{\beta}_j$ for $\gamma_j \in \{0,1\}$ are given by*

$$CPONZ_\xi = f(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2|\gamma_j = 1, \sigma^2)\frac{f(\sigma^2|m_1)}{f(\sigma^2|m_0)}\frac{f(m_1)}{f(m_0)} \tag{6.21}$$

Proof: See appendix, page 204. $\triangleleft$

Robert (1993) holds $CPOZ_\xi$ constant for a specific ξ. George and McCulloch (1993) adopt a similar methodology in defining the prior distributions in semiautomatic prior selection of SSVS. We further propose a simpler approach based on the 'conditional prior odds at zero' (CPOZ) which is given by the following definition.

Definition 6.3 *The conditional prior odds at zero for comparing two normal linear models $\boldsymbol{y} \sim N(\boldsymbol{\mu}_{(m_{\gamma_j})}, \sigma^2\boldsymbol{I}_n)$ with $\boldsymbol{\mu}_{(m_{\gamma_j})} = \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma_j\boldsymbol{X}_j\boldsymbol{\beta}_j$ for $\gamma_j \in \{0,1\}$ is defined as*

$$CPOZ = \frac{f(m_1 \mid \boldsymbol{\beta}_j = 0, \sigma^2)}{f(m_0 \mid \boldsymbol{\beta}_j = 0, \sigma^2)}. \tag{6.22}$$

Proposition 6.6 *The conditional prior odds at zero of the definition 6.3 for comparing two normal linear models $\boldsymbol{y} \sim N(\boldsymbol{\mu}_{(m_{\gamma_j})}, \sigma^2\boldsymbol{I}_n)$ with $\boldsymbol{\mu}_{(m_{\gamma_j})} = \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma_j\boldsymbol{X}_j\boldsymbol{\beta}_j$ for $\gamma_j \in \{0,1\}$ are given by*

$$CPOZ = \frac{f(m_1|\ \boldsymbol{\beta}_j = \boldsymbol{0}, \sigma^2)}{f(m_0|\boldsymbol{\beta}_j = \boldsymbol{0}, \sigma^2)} = f(\boldsymbol{\beta}_j = \boldsymbol{0}|\sigma^2, \gamma_j = 1)\frac{f(\sigma^2|m_1)}{f(\sigma^2|m_0)}\frac{f(m_1)}{f(m_0)}. \tag{6.23}$$

Proof: See appendix, page 204. $\triangleleft$

From the above proposition it is clear that when the prior variance of $\boldsymbol{\beta}_j$ increases then the conditional density $f(\boldsymbol{\beta}_j = \boldsymbol{0}|\gamma_j = 1)$ decreases and hence the conditional prior odds support more strongly the simplest model. Under this view it is natural that the posterior odds will also tend to support the simplest model.

If we consider the independent normal prior setup (6.12) then we have that the conditional prior odds at zero are equal to

$$CPOZ = \frac{f(m_1|\ \boldsymbol{\beta}_j = \boldsymbol{0}, \sigma^2)}{f(m_0|\boldsymbol{\beta}_j = \boldsymbol{0}, \sigma^2)} = \frac{f(m_1)}{f(m_0)}(2\pi c^2\sigma^2)^{-d_j/2}|\boldsymbol{X}_j^T\boldsymbol{X}_j|^{1/2} \tag{6.24}$$

which demonstrates the same type of incoherency for large values of c^2.

In the cases of non-nested models we may adopt the following more general definition of conditional prior odds.

Definition 6.4 *The conditional prior odds at zero for comparing two linear normal models m_0 and m_1 with $\boldsymbol{y} \sim N(\boldsymbol{\mu}_{(m_\gamma)}, \sigma^2 \boldsymbol{I}_n)$ where $\boldsymbol{\mu}_{(m_\gamma)} = (1-\gamma)\boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma \boldsymbol{X}_{(m_1)}\boldsymbol{\beta}_{(m_1)}$ for $\gamma \in \{0, 1\}$ is defined as*

$$CPOZ = \frac{f(m_1 \mid \boldsymbol{\beta}_{(m_0)} = \boldsymbol{0}, \boldsymbol{\beta}_{(m_1)} = \boldsymbol{0}, \sigma^2)}{f(m_0 \mid \boldsymbol{\beta}_{(m_0)} = \boldsymbol{0}, \boldsymbol{\beta}_{(m_1)} = \boldsymbol{0}, \sigma^2)}. \quad (6.25)$$

Proposition 6.7 *The conditional prior odds at zero of the definition 6.4 for comparing two linear normal models m_0 and m_1 with $\boldsymbol{y} \sim N(\boldsymbol{\mu}_{(m_\gamma)}, \sigma^2 \boldsymbol{I}_n)$ where $\boldsymbol{\mu}_{(m_\gamma)} = (1-\gamma)\boldsymbol{X}_{(m_0)}\boldsymbol{\beta}_{(m_0)} + \gamma \boldsymbol{X}_{(m_1)}\boldsymbol{\beta}_{(m_1)}$ for $\gamma \in \{0, 1\}$ are given by*

$$CPOZ = \frac{f(\boldsymbol{\beta}_{(m_1)} = \boldsymbol{0} | \sigma^2, m_1)}{f(\boldsymbol{\beta}_{(m_0)} = \boldsymbol{0} | \sigma^2, m_0)} \frac{f(\sigma^2 | m_1)}{f(\sigma^2 | m_0)} \frac{f(m_1)}{f(m_0)}. \quad (6.26)$$

Proof: See appendix, page 204. $\triangleleft$

The problem of incoherency is not directly evident in the above quantity but if we consider the general prior setup of Section 6.2.1 then we have

$$\begin{aligned} CPOZ = & \frac{f(m_1)}{f(m_0)} (2\pi c^2 \sigma^2)^{-(d(m_1)-d(m_0))/2} \left(\frac{|\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{V}_{(m_0)}|} \right)^{-1/2} \\ \times & \; exp\left(-\frac{1}{2c^2\sigma^2} \left\{ \boldsymbol{\mu}^T_{\beta_{(m_1)}} \boldsymbol{V}^{-1}_{(m_1)} \boldsymbol{\mu}_{\beta_{(m_1)}} - \boldsymbol{\mu}^T_{\beta_{(m_0)}} \boldsymbol{V}^{-1}_{(m_0)} \boldsymbol{\mu}_{\beta_{(m_0)}} \right\} \right) \end{aligned}$$

in which we clearly see that, for fixed matrices $\boldsymbol{V}_{(m)}$, as c^2 gets larger then the above conditional prior odds fully support the simpler model. Similar results are obtained if we use the prior setup of (6.10) resulting in

$$CPOZ = \frac{f(m_1)}{f(m_0)} (2\pi c^2 \sigma^2)^{-(d(m_1)-d(m_0))/2} \left(\frac{|\boldsymbol{X}^T_{(m_1)} \boldsymbol{X}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_{(m_0)}|} \right)^{1/2}$$

which clearly depends on the magnitude of c^2.

Finally, if we consider the independent normal prior setup (6.12) then both definitions 6.3 and 6.4 (for nested and non-nested models) give the same conditional prior odds at zero (equation 6.24).

6.4 Prior Specification via Penalty Determination

6.4.1 Prior Odds and Penalty Specification

A different approach is proposed here following the logic that prior odds and prior variance of model parameters should depend on each other. From previous sections we clearly saw that different prior distributions (indirectly) impose different penalty to the log-likelihood ratio. This penalty depends on three elements:

1. the magnitude of the prior variance expressed by c^2 (see equation 6.9),
2. the logarithm of the determinants ratio of prior and posterior covariance structure of the two models (see equation 6.9) and
3. the prior odds (see equation 6.9).

We propose a calibrating method that will enable us to use a prior distribution as non-informative as we would like, within each model, and, at the same time, apply the dimensionality penalty we wish. We limit ourselves in the normal cases described above but generalization to other models can be done using Laplace approximation; see Section 6.6. As we have already mentioned the posterior odds can be written as model selection criteria given by (6.7). The penalty function ψ can be further written as

$$\psi = \psi' - 2log\left(\frac{f(m_1)}{f(m_0)}\right)$$

where ψ' is the penalty function when equal prior model probabilities are considered. In such cases, instead of using the uniform prior distribution on model space we may express our 'model selection' opinion or prior information in terms of penalty for each additional parameter used. Therefore, if we want to assign penalty $F = log(\kappa)$ for each additional parameter we would simply use prior probabilities that satisfy the equality

$$\frac{f(m_1)}{f(m_0)} = exp\left(\frac{1}{2}\{\psi' - [d(m_1) - d(m_0)]F\}\right) = e^{\psi'/2}/\kappa^{[d(m_1)-d(m_0)]/2}.$$

Moreover, if the penalty ψ' can be written as a difference between two penalties $\psi' = \psi'_{m_1} - \psi'_{m_0}$ attributed to each model then we can simply set

$$f(m) \propto exp\left(\frac{1}{2}\{\psi'_m - d(m)F\}\right) = e^{\psi'_m/2}/\kappa^{d(m)/2}.$$

In the cases examined (general multivariate normal prior distribution and independent normal distributions), both c^2 and the prior odds have great effect on the penalty imposed on the likelihood. Our aim is to use 'non-informative' priors on model parameters without being informative in the model selection process. This can be achieved by suitably specifying the prior probabilities which control the model selection procedure with the desired penalty imposed by the user. Similar ideas were used by Poskitt and Tremayne (1983) for Jeffreys prior and Pericchi (1984). A suitable specification of prior model probabilities via penalty determination leads to the following propositions.

Proposition 6.8 *Consider two nested normal linear models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n\right)$$

and prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\mathbf{0}, c^2\boldsymbol{V}_{(m)}\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$. When we use prior probabilities given by

$$f(m) \propto \left(\frac{c^2}{\kappa}\right)^{d(m)/2} \left|\boldsymbol{V}_{(m)}\right|^{1/2} \left|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{V}^{-1}_{(m)}\right|^{1/2} \tag{6.27}$$

then, for large c^2, $-2log(PO_{01})$ is approximately equal to an information criterion of type (6.5) with penalty function

$$\psi = \{d(m_1) - d(m_0)\}log(\kappa),$$

or penalty for each additional parameter equal to $F = log(\kappa)$.

Proof: See appendix, page 205. ◃

Corollary 6.8.1 *Consider two normal linear models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2\boldsymbol{I}_n\right)$$

and prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\mathbf{0}, c^2\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\right)^{-1}\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$. When we use prior probabilities

$$f(m) \propto \left(\frac{c^2+1}{\kappa}\right)^{d(m)/2} \tag{6.28}$$

then, for large c^2, $-2log(PO_{01})$ is equal to a criterion of type (6.5) with penalty function

$$\psi = \{d(m_1) - d(m_0)\}log(\kappa),$$

or penalty for each additional parameter equal to $F = log(\kappa)$.

Proof: See appendix, page 205. $\triangleleft$

The parameter κ now fully controls the penalty that is imposed on the log-likelihood. We can easily adopt this kind of prior in variable selection problems using binary indicator variables $\boldsymbol{\gamma}$ as in George and McCulloch (1993). When no restrictions on the model space are imposed, a common prior for each term probability is given by $\gamma_j \sim Bernoulli(\pi)$ which can be written as

$$f(\boldsymbol{\gamma}) \propto \left(\frac{\pi}{1-\pi}\right)^{d(\boldsymbol{\gamma})} = [PrO]^{d(\boldsymbol{\gamma})} \tag{6.29}$$

denoting that the prior probability of a model depends on its dimension and parameter PrO which measures the prior odds of including any term in the model equation. A common choice, considered as non-informative prior, is given for $PrO = 1$.

Corollary 6.8.2 *Consider the model comparison of proposition 6.8.1. If we use the variable selection approach using the latent variables $\boldsymbol{\gamma}$ instead of the model indicator m with prior*

$$\gamma_j \sim Bernoulli(\pi_j)$$

and

$$PrO_j = \frac{f(\gamma_j = 1)}{f(\gamma_j = 0)} = \frac{\pi_j}{1-\pi_j} = \left(\frac{c^2+1}{\kappa}\right)^{d_j/2}$$

(where d_j are the number of parameters for the j term) then, for large c^2, $-2log(PO_{01})$ is equal to a criterion of type (6.5) with penalty function

$$\psi = \{d(m_1) - d(m_0)\}log(\kappa)$$

or penalty for each additional parameter equal to $F = log(\kappa)$.

Proof: See appendix, page 205. $\triangleleft$

Corollary 6.8.3 *Consider two normal models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ *and additionally that*

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j]$$

with independent normal prior distributions that can be summarised by the prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\mathbf{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}^T_j \boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2},$$

and prior probabilities

$$f(m) \propto \left(\frac{c^2}{\kappa}\right)^{d(m)/2} \left|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}D_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\right|^{1/2} \prod_{j=1}^{p} |\boldsymbol{X}^T_j\boldsymbol{X}_j|^{-\gamma_{j,m}/2}. \quad (6.30)$$

Then the resulted penalty is given by

$$\psi = \{d(m_1) - d(m_0)\} log(\kappa),$$

or penalty for each additional parameter equal to $F = log(\kappa)$.

Proof: See appendix, page 205. ◁

6.4.2 Conditional Prior Odds Using Penalty Determination

The above proposed method is coherent in terms of conditional posterior odds at zero since this measure remains constant as c^2 increases. The following propositions give the form of conditional prior odds at zero when prior model probabilities are constructed via penalty specification. Their behaviour when c^2 is large, and hence when low prior information within each model is used, is discussed in detail.

Proposition 6.9 *Consider two normal linear models* m_0 *and* m_1 *which are given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n\right),$$

prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\mathbf{0}, c^2 \boldsymbol{V}_{(m)}\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$, *and prior model probabilities given by*

$$f(m) \propto \left(\frac{c^2}{\kappa}\right)^{d(m)/2} \left|\boldsymbol{V}_{(m)}\right|^{1/2} \left|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{V}^{-1}_{(m)}\right|^{1/2}.$$

Then the prior odds at zero are given by

$$
\begin{aligned}
CPOZ = & \quad (2\pi\kappa\sigma^2)^{-(d(m_1)-d(m_0))/2} \left(\frac{\left|\boldsymbol{X}_{(m_1)}^T \boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}_{(m_1)}^{-1}\right|}{\left|\boldsymbol{X}_{(m_0)}^T \boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}_{(m_0)}^{-1}\right|} \right)^{1/2} \\
& \times \; exp\left(-\frac{1}{2c^2\sigma^2} \left\{ \boldsymbol{\mu}_{\beta_{(m_1)}}^T \boldsymbol{V}_{(m_1)}^{-1} \boldsymbol{\mu}_{\beta_{(m_1)}} - \boldsymbol{\mu}_{\beta_{(m_0)}}^T \boldsymbol{V}_{(m_0)}^{-1} \boldsymbol{\mu}_{\beta_{(m_0)}} \right\} \right)
\end{aligned}
$$

Proof: See appendix, page 206. $\triangleleft$

When we use the above prior setup and c^2 is very large (and therefore non-informative within each model) then

$$
CPOZ \approx (2\pi\kappa\sigma^2)^{-(d(m_1)-d(m_0))/2} \left(\frac{\left|\boldsymbol{X}_{(m_1)}^T \boldsymbol{X}_{(m_1)}\right|}{\left|\boldsymbol{X}_{(m_0)}^T \boldsymbol{X}_{(m_0)}\right|} \right)^{1/2} < \infty,
$$

that is the conditional prior odds do not fully support the simpler model and therefore the above prior setup is coherent and robust for large values of c^2.

Corollary 6.9.1 *Consider two normal linear models m_0 and m_1 given by*

$$
\boldsymbol{y} \sim N\left(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n\right)
$$

and prior setup

$$
f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 \left(\boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}\right)^{-1} \sigma^2\right) \;\; \text{and } f(\sigma^2) \propto \sigma^{-2}
$$

for both $m \in \{m_0, m_1\}$. When we use prior probabilities

$$
f(m) \propto \left(\frac{c^2+1}{\kappa}\right)^{d(m)/2}
$$

then the prior odds at zero are given by

$$
CPOZ = (2\pi\kappa\sigma^2)^{-(d(m_1)-d(m_0))/2} \left(\frac{c^2}{c^2+1}\right)^{-[d(m_1)-d(m_0)]/2} \left(\frac{\left|\boldsymbol{X}_{(m_1)}^T \boldsymbol{X}_{(m_1)}\right|}{\left|\boldsymbol{X}_{(m_0)}^T \boldsymbol{X}_{(m_0)}\right|} \right)^{1/2}.
$$

Proof: See appendix, page 206. $\triangleleft$

Corollary 6.9.2 *Consider two normal models m_0 and m_1 given by*

$$
\boldsymbol{y} \sim N(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n)
$$

for both $m \in \{m_0, m_1\}$ and additionally that

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j]$$

with independent normal prior distributions that can be summarised by the prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\mathbf{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}^T_j \boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

and prior probabilities

$$f(m) \propto \left(\frac{c^2}{\kappa}\right)^{d(m)/2} \left|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}D_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\right|^{1/2} \prod_{j=1}^{p} |\boldsymbol{X}^T_j\boldsymbol{X}_j|^{-\gamma_{j,m}/2}$$

then the prior odds at zero are equal to

$$CPOZ = (2\pi\kappa\sigma^2)^{-(d(m_1)-d(m_0))/2}\left(\frac{\left|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}D_{(m_1)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\right|}{\left|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\right|}\right)^{1/2}.$$

Proof: See appendix, page 206. $\triangleleft$

From the above we may conclude that using the prior probabilities proposed in Section 6.4.1 then the 'conditional prior odds at zero' do not depend on variance multiplicator c^2. Moreover, we can use this quantity to define the appropriate a-priori penalty which, according to our prior belief, is correct. For example, a quick approach can be based on defining the conditional prior odds at zero such that strongly support the simpler model. Therefore, defining $CPOZ = 1/99$ for nested models is equivalent to giving a-priori a probability 1% to the simplest model when the model parameter vector is equal to zero. The more we support the simplest model via $CPOZ$ the more penalty is applied to the log-ratio of posterior sum of squares.

6.5 Posterior Odds at the Limit of Significance

In this section we use the posterior odds at the limit of significance and the logic of Lindley (1957) in order to examine its behaviour in normal linear models and associate classical and Bayesian approach in hypothesis tests and model selection. The section is divided into three sub-sections. We firstly define the posterior odds at the limit of significance and we present its behaviour in both Lindley's example and normal linear model. Then, we consider their robustness after eliminating the variance effect and finally we present an alternative methodology which equates posterior probability of the null hypothesis (or model) to a prespecified significance level q.

6.5.1 Posterior Odds at the Limit of Significance and Lindley's Example

Lindley (1957) used the simple example where $y \sim N(\mu, \sigma^2)$, with σ^2 known, to test the hypothesis that $H_0 : \mu = \mu_0$ *vs.* $H_1 : \mu \neq \mu_0$. We will alternatively consider a normal prior distribution with mean μ_0 and variance $c^2\sigma^2$ in order to have similar patterns with the previous sections.

Definition 6.5 *The 'posterior odds at the limit of significance' (POLS) are defined as the posterior odds resulting from observed samples that are in the limit of rejection area of a significance test of level q.*

For Lindley's example such samples satisfy the equality $\bar{y} = \mu_0 \pm z_{q/2}\sigma/\sqrt{n}$.

Proposition 6.10 *Consider the simple case with $y \sim N(\mu, \sigma^2)$ where σ^2 is known. We want to assess which of the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ is supported by the data (and prior beliefs) using the prior $f(\mu|\sigma^2) \sim N(\mu_0, c^2\sigma^2)$. The posterior odds at the limit of significance are then given by*

$$POLS_{01}^{q} = \frac{f(m_0)}{1 - f(m_0)} \sqrt{nc^2 + 1}\; exp\left(-\frac{1}{2} \frac{nc^2}{nc^2 + 1} z_{q/2}^2 \right) \tag{6.31}$$

where z_q is the q quantile of the standardised normal distribution.

Proof: See appendix, page 206. $\triangleleft$

The paradox noted by Lindley (1957) also holds for this alternative prior and therefore for $n \to \infty$, $POLS_{01}^{q} \to \infty$. Similarly for fixed sample size when $c^2 \to \infty$, then $POLS_{01}^{q} \to \infty$. Figure 6.1 shows how log-posterior odds at the limit of 5% significance, increase for sample sizes from one to 100 when $\mathcal{M} = \{m_0, m_1\}$ and $f(m_0) = f(m_1) = 1/2$ for various values of c, while Figure 6.2 gives the corresponding plot for the posterior probability $f(m_0|\boldsymbol{y})$.

Similarly, we can define the posterior odds at the limit of significance for normal linear and generalised linear models using either the F-distribution or the χ^2 distribution respectively.

We will briefly examine the two special model comparisons presented in propositions 6.2 and 6.3 and the behaviour of these posterior odds at the limit of significance.

Let us consider the general case of Section 6.2.1 in which the two models under comparison differ only in j term. In such case and for large c^2 we have from proposition 6.1 that the posterior odds at the limit of significance are given by

$$POLS_{01}^{q} \approx c^{d(m_1)-d(m_0)} \left(\frac{|\boldsymbol{V}_{(m_0)}|}{|\boldsymbol{V}_{(m_1)}|} \right)^{-1/2} \left(\frac{|\tilde{\boldsymbol{\Sigma}}_{(m_0)}|}{|\tilde{\boldsymbol{\Sigma}}_{(m_1)}|} \right)^{1/2} \frac{f(m_0)}{f(m_1)} \left(1 + \frac{d_j}{n - d(m_1)} F_{d_j, n-d(m_1), 1-q} \right)^{-n/2}$$

where $F_{n_1,n_2,q}$ is the qth quantile of F distribution with degrees of freedom n_1 and n_2; see also Spiegelhalter and Smith (1982). If we further consider the prior setup (6.12) the above quantity simplifies to

$$POLS_{01}^{q} \approx (c^2+1)^{d_j/2} |\boldsymbol{X}_j^T \boldsymbol{X}_j|^{-1/2} |\boldsymbol{X}_j^T \boldsymbol{\Delta}_{(m_0)} \boldsymbol{X}_j|^{1/2} \frac{f(m_0)}{f(m_1)} \left(1 + \frac{d_j}{n - d(m_1)} F_{d_j, n-d(m_1), 1-q} \right)^{-n/2}. \tag{6.32}$$

6.5.2 Posterior Odds at the Limit of Significance and Prior Specification Using Penalty Determination

In following section we present the simple Lindley's example and the general model comparison of two nested models using prior odds proposed in Section 6.4.1. This prior specification leads to the following two propositions.

Proposition 6.11 *Consider the simple case with $y \sim N(\mu, \sigma^2)$ where σ^2 is known and we want to assess which of the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ is supported by the data (and prior beliefs) using the prior $f(\mu|\sigma^2) \sim N(\mu_0, c^2\sigma^2)$ and prior probability for the null model $f(m_0) = 1/(1+c)$. The posterior odds at the limit of significance are then, for large c^2, given by*

$$POLS_{01}^{q} \approx \sqrt{n}\ exp\left(-\frac{1}{2} z_{q/2}^2 \right). \tag{6.33}$$

Proof: See appendix, page 206. ◁

Similar arguments can be used in nested model comparisons of normal models leading to the following proposition.

Proposition 6.12 *Consider two normal models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ and additionally that

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j]$$

with independent normal prior distributions that can be summarised by the prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\mathbf{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

and prior probabilities

$$f(m) \propto \left(\frac{c^2}{\kappa}\right)^{d(m)/2} \left|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}D_{(m)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right|^{1/2} \prod_{j=1}^{p} |\boldsymbol{X}_j^T\boldsymbol{X}_j|^{-\gamma_{j,m}/2}.$$

Then the posterior odds at the limit of significance are given by

$$POLS^q_{01} \approx \kappa^{d_j/2}\left(1 + \frac{d_j}{n-d(m_1)}F_{d_j, n-d(m_1), 1-q}\right)^{-n/2}.$$

Proof: See appendix, page 207. ◁

From the previous two propositions it is evident that the dimensionality adjustment of prior odds proposed in Section 6.4.1 eliminates the effect of the prior variance on the posterior odds at the limit of significance. Figures 6.3 and 6.4 show clearly that log-posterior odds and posterior probabilities are now sensitive only to the sample size n.

6.5.3 Specification of Prior Distributions Using P-values

An alternative method for the specification of the prior model probabilities can be adopted using the relation of posterior odds at the limit of significance and p-values. For example, we may want to set the posterior probability at the limit of significance equal to q and therefore the posterior odds at the limit of significance equal to $q/(1-q)$. This approach is briefly presented here for Lindley's example and the nested model comparison of normal models.

Proposition 6.13 *Consider the simple case with $y \sim N(\mu, \sigma^2)$ where σ^2 is known and we want to assess which of the hypotheses $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ is supported by the data* *(and prior beliefs) using the prior $f(\mu|\sigma^2) \sim N(\mu_0, c^2\sigma^2)$ and prior probability for the null model*

$$f(m_0) = \frac{1}{1 + \sqrt{n} c exp\left(-\frac{1}{2} z^2_{q/2}\right)(1-q)/q}. \tag{6.34}$$

The posterior odds at the limit of significance are then, for large c^2, given by

$$POLS^q_{01} \approx \frac{q}{1-q}$$

and the corresponding posterior probability is equal to the significance level q, and therefore we have $f(m_0|\boldsymbol{y}) = q$.

Proof: See appendix, page 207. $\triangleleft$

Proposition 6.13 shows that, for this simple example, we can find prior odds that will result to posterior probability for samples at the limit of $100q\%$ significance level approximately equal to q. Figures 6.5 and 6.6 show clearly that log-posterior odds and posterior probabilities are robust to various choices of the prior parameter c^2 and the sample size n.

Similar arguments can be used for unknown σ^2 using the improper prior on the precision $f(\tau) \propto [\tau]^{-1}$ with $\tau = \sigma^{-2}$, resulting to posterior odds

$$PO_{01} = \sqrt{nc^2+1}\left(\frac{(n-1)s^2 + \frac{n}{nc^2+1}(\bar{y}-\mu_0)^2}{(n-1)s^2 + n(\bar{y}-\mu_0)^2}\right)^{n/2}$$

where s^2 is the unbiased variance estimator. Using now the significance test based on the Student distribution the limit is given by samples with $\bar{y} = \mu_0 + t_{n,q/2}s/\sqrt{n}$; where $t_{n,q}$ is the q quintile of the Student distribution with n degrees of freedom. The resulting posterior odds are given by

$$POLS^q_{01} = \sqrt{nc^2+1}\left(\frac{(n-1) + \frac{1}{nc^2+1}t^2_{n,q/2}}{(n-1) + t^2_{n,q/2}}\right)^{n/2}.$$

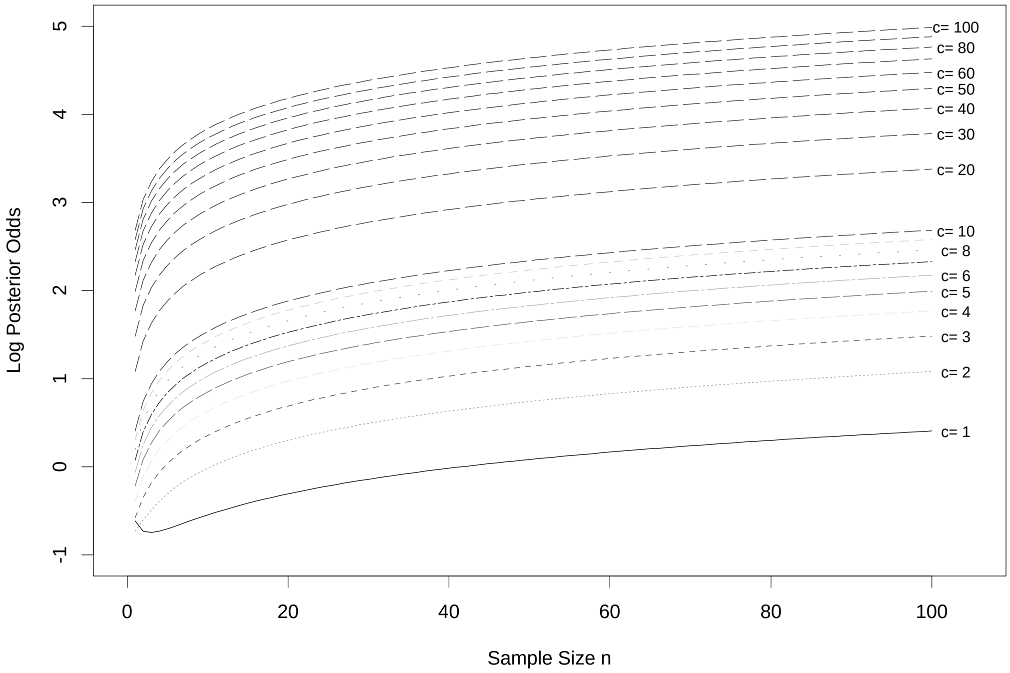


Figure 6.1: Logarithm of Posterior Odds of $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ at the Limit of 5% Significance with Prior Odds Equal to 1.

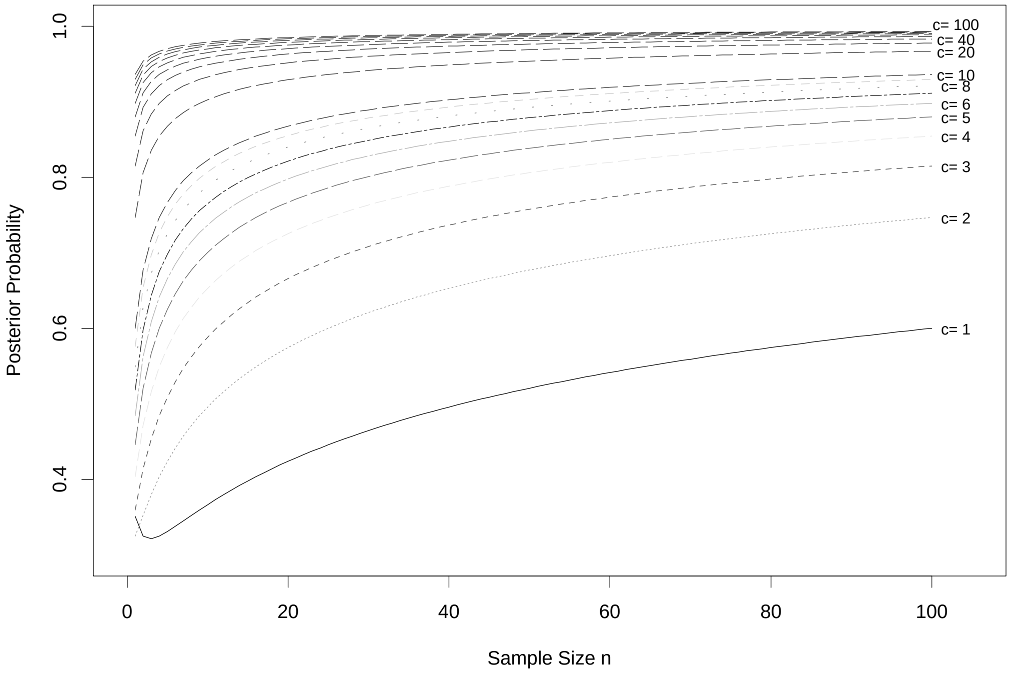


Figure 6.2: Posterior Probabilities of Hypothesis $H_0 : \mu = 0$ (vs. $H_1 : \mu \neq 0$) at the Limit of 5% Significance with Prior Odds Equal to 1.

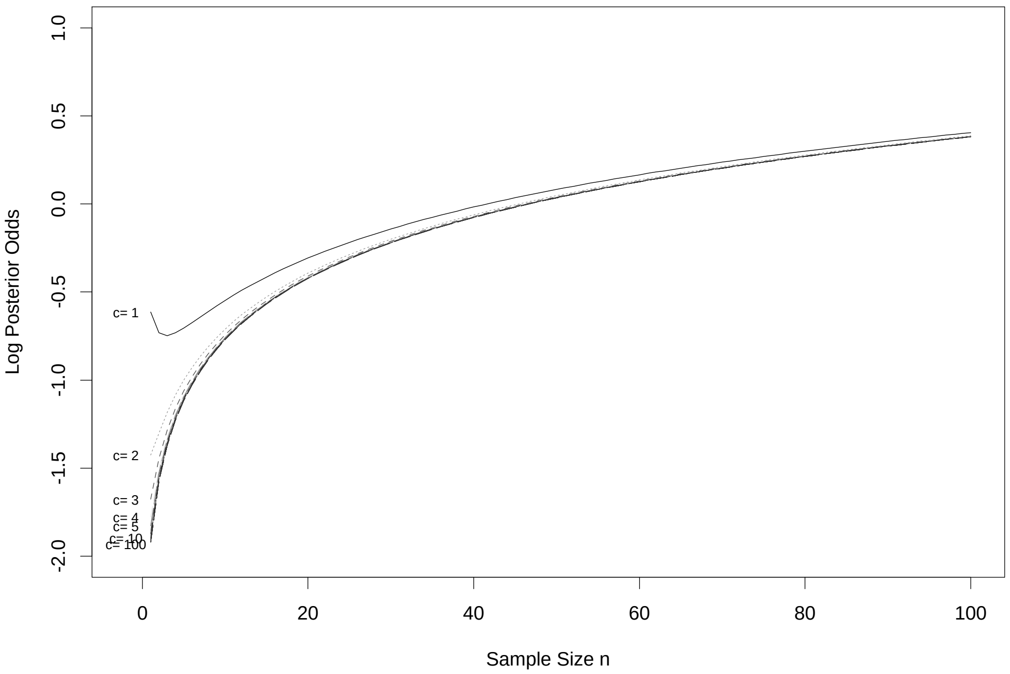


Figure 6.3: Logarithm of Posterior Odds of $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ at the Limit of 5% Significance with Prior Odds Equal to $1/c$.

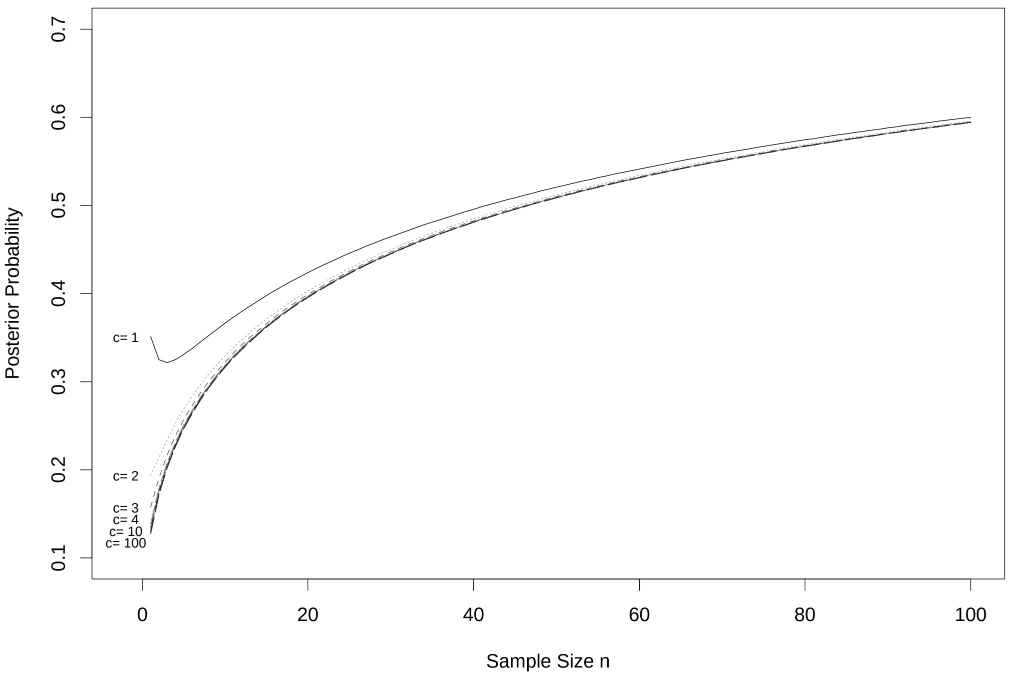


Figure 6.4: Posterior Probabilities of Hypothesis $H_0 : \mu = 0$ (vs. $H_1 : \mu \neq 0$) at the Limit of 5% Significance with Prior Odds Equal to $1/c$.

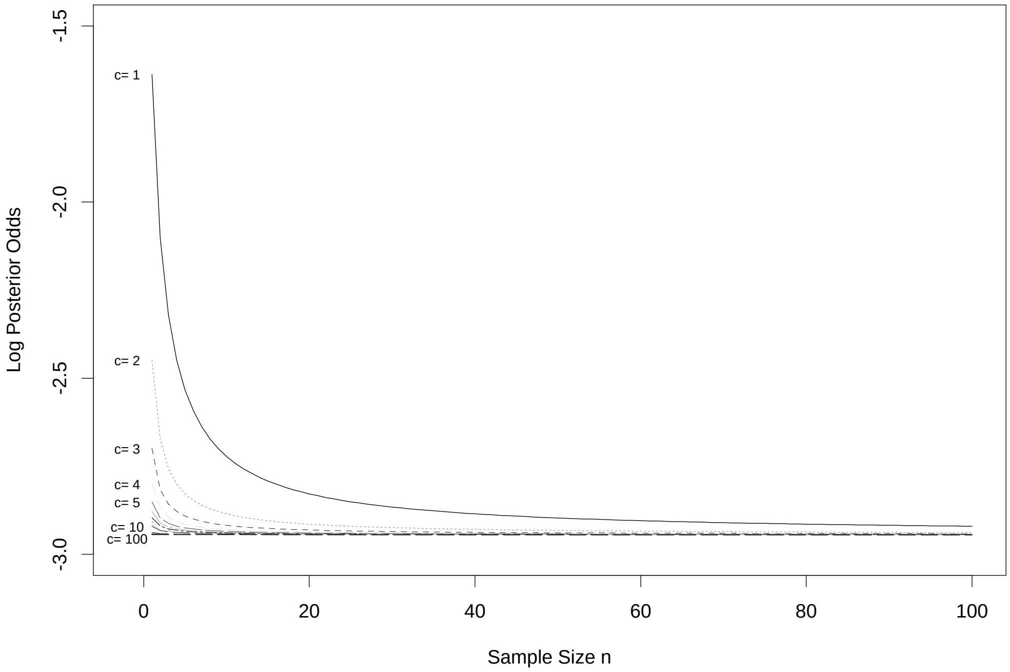


Figure 6.5: Logarithm of Posterior Odds of $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$ at the Limit of 5% Significance Using Prior Probability which Eliminates Both Prior Variance and Sample Size Effect.

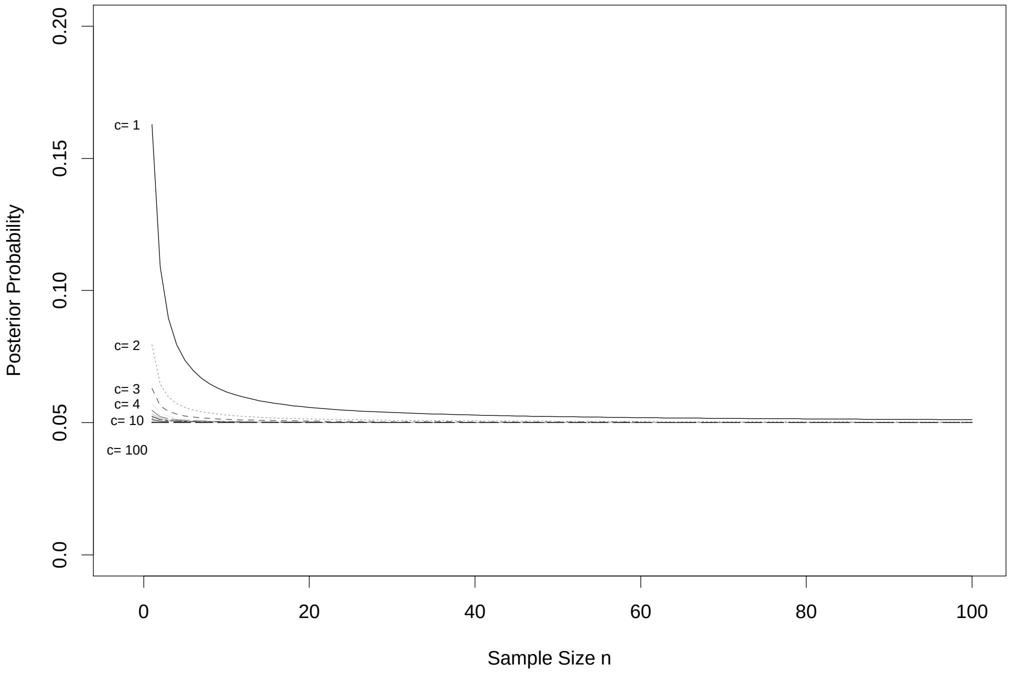


Figure 6.6: Posterior Probabilities of Hypothesis $H_0 : \mu = 0$ (vs. $H_1 : \mu \neq 0$) at the Limit of 5% Significance Using Prior Probability which Eliminates Both Prior Variance and Sample Size Effect.

Consider now the model comparison of two nested models, then we have the following proposition.

Proposition 6.14 *Consider two normal nested models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N(\boldsymbol{X}_{(m)}\boldsymbol{\beta}_{(m)}, \sigma^2 \boldsymbol{I}_n)$$

for both $m \in \{m_0, m_1\}$ and additionally that

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^T_{(m_0)}, \boldsymbol{\beta}^T_j]$$

with independent normal prior distributions that can be summarised by the prior setup

$$f(\boldsymbol{\beta}_{(m)}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 D^{-1}_{(m)}(\boldsymbol{X}_j^T \boldsymbol{X}_j)\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

and prior model probabilities

$$\frac{f(m_1)}{f(m_0)} = (c^2+1)^{d_j/2}\frac{1-q}{q}|\boldsymbol{X}_j^T\boldsymbol{X}_j|^{-1/2}|\boldsymbol{X}_j^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j|^{1/2}\left(1+\frac{d_j}{n-d(m_1)}F_{d_j,n-d(m_1),1-q}\right)^{-n/2}. \tag{6.35}$$

Then the posterior odds at the limit of significance are given by

$$POLS^q_{01} \approx \frac{q}{1-q}.$$

Proof: See appendix, page 207. ◁

When we deal with more than two models, we may compare all models with a baseline (e.g. either the full or the null) model and specify all these posterior odds at the limit of significance to be equal to $q/(1-q)$.

Under the p-value prior specification approach presented in proposition 6.14, the prior odds at zero are given by

$$\begin{aligned} CPOZ = & \left(\frac{c^2+1}{c^2}\right)^{d_j/2}\frac{1-q}{q}(2\pi\sigma^2)^{-d_j/2}|\boldsymbol{X}_j^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j|^{1/2} \\ \times & \left(1+\frac{d_j}{n-d(m_0)-d_j}F_{d_j,n-d(m_0)-d_j,1-q}\right)^{-n/2}. \end{aligned}$$

The above prior odds, for large c^2, tends to a given quantity for finite and fixed sample size n and is consistent in terms of the behaviour of prior odds at zero when low information, within each model, is used.

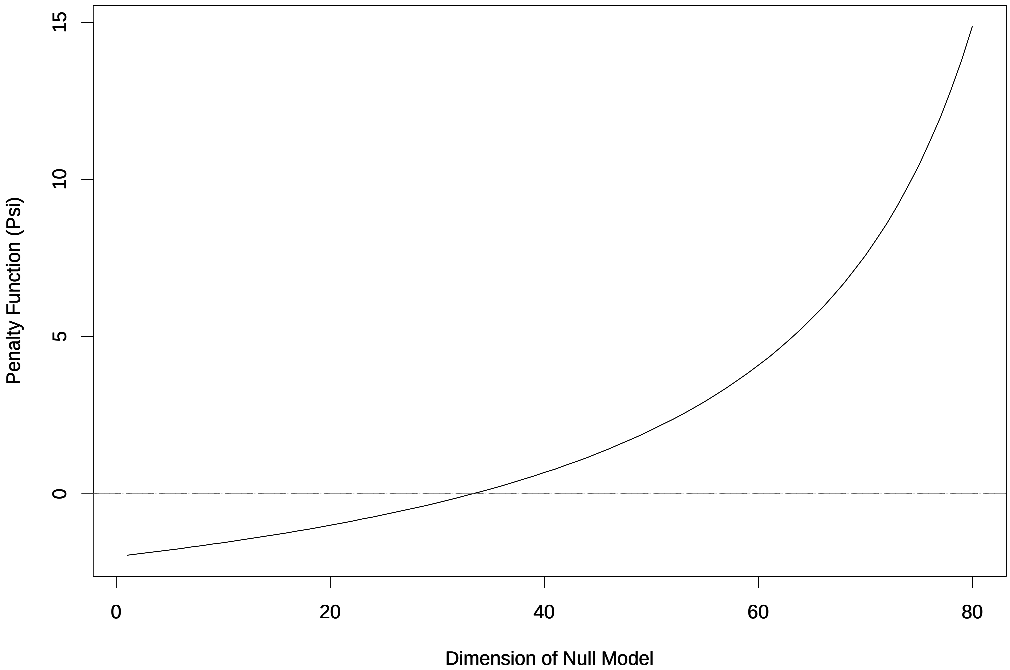


Figure 6.7: Plot of Penalty Function Against the Dimension of the Null Model When the Posterior Probability at the Limit of Significance is Fixed at 5% for 100 Observations.

Although the prior odds at zero are well behaved for large values of c^2, the penalty imposed on the logarithm of the posterior demonstrates the following implausible behaviour for large sample sizes. Under the prior setup of proposition 6.14, the imposed penalty is given by

$$\psi = n log \left(1 + \frac{d_j}{n - d(m_0) - d_j} F_{d_j, n-d(m_0)-d_j, 1-q} \right) - 2 log \left(\frac{1-q}{q} \right).$$

When n tends to infinity and for fixed dimensionality difference, the above penalty becomes

$$\begin{aligned} \lim_{n \to \infty} \psi &= \lim_{n \to \infty} log \left(1 + \frac{d_j}{n - d(m_0) - d_j} F_{d_j, n-d(m_0)-d_j, 1-q} \right)^n - 2 log \left(\frac{1-q}{q} \right) && (6.36) \\ &= d_j F_{d_j, \infty, 1-q} - 2 log \left(\frac{1-q}{q} \right) && (6.37) \\ &= \chi^2_{d_j, 1-q} - 2 log \left(\frac{1-q}{q} \right). && (6.38) \end{aligned}$$

Such a penalty is incoherent because, for large n and small d_j, it is not only very small but, in some extreme cases, also negative. The negative penalty supports more complicated models and therefore the above expression also demonstrates why significance tests should be avoided when the sample size is large; for example Figure 6.7 demonstrates how the penalty function changes with the dimension of the simpler model, $d(m_0)$, when the model dimension difference d_j is equal to one and the sample size n is equal to 100 (the first 33 values of the penalty function are negative).

6.6 Prior Specification via Penalty Determination in Generalised Linear Models

The primary purpose of this section is to extend the methodology presented in the previous sections to the generalised linear model using the Laplace approximation. Two sub-sections are presented. Firstly, a model selection criterion is presented as well as a different approach based on the posterior distribution of imaginary prior data points. Secondly, construction of prior distributions via specification of the penalty function in generalised linear models is presented in detail.

6.6.1 Posterior Odds, Maximum Likelihood Ratios and Information Criteria Using Laplace Approximation

If we use Laplace approximation it is straightforward to express the posterior odds in an criterion form given by

$$-2log(PO_{01}) \approx -2log\left(\frac{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_0)}, m_0) f_N(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)}{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_1)}, m_1) f_N(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)}\right) - \psi$$

$$\psi = 2log\left(\frac{C[f(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)]}{C[f(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)]}\right) - log\left(\frac{|\mathcal{I}_{\breve{\boldsymbol{\beta}}_{(m_1)}}|}{|\mathcal{I}_{\breve{\boldsymbol{\beta}}_{(m_0)}}|}\right) - [d(m_1) - d(m_0)]log(2\pi) - 2log\left(\frac{f(m_1)}{f(m_0)}\right), \tag{6.39}$$

where $\breve{\boldsymbol{\beta}}_{(m)}$ is the posterior mode and $f_N(x)$ is the non-normalised density function of $f(x)$ and $C[f(x)]$ is the corresponding normalising constant, [that is $f(x) = f_N(x)/C[f(x)] \propto f_N(x)$] and

$$\mathcal{I}_{\breve{\boldsymbol{\beta}}_{(m)}} = -\left[\frac{\partial^2 f_{m,y}(\boldsymbol{\beta}_{(m)})}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\breve{\boldsymbol{\beta}}_{(m)}},$$

$$f_{m,y}\left(\boldsymbol{\beta}_{(m)}\right) = log\left[f\left(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m\right)\right] + log\left[f\left(\boldsymbol{\beta}_{(m)}|m\right)\right].$$

The ratio appearing in the above expression is not natural. We will try to interpret the above result as functions of maximum likelihood ratios and therefore argue why such an expression should used in generalised linear models instead of the simple log-likelihood ratio in classical information criteria or the log-ratio of posterior sum of squares in the model selection criteria presented for the normal linear case.

Consider an approach similar to Chen *et al.* (1999). We assume that instead of using an arbitrary prior distribution, we have $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$ prior points expressing our prior opinion. We adopt the prior

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{X}^*_{(m)}, \boldsymbol{y}^*) \propto f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}^*_{(m)}, m).$$

Then we have the following proposition.

Proposition 6.15 *The posterior odds of model m_0 against model m_1 using prior of the form*

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, m) \propto f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}^*_{(m)}, m)$$

where $\boldsymbol{y}^*$ are imaginary data that express our prior information, can be written in a form equivalent to information criteria given by

$$-2log(PO_{01}) \approx -2log\left(\frac{f(\boldsymbol{y},\boldsymbol{y}^*|\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m_0)}, m_0)/f(\boldsymbol{y}^*|\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m_0)}, m_0)}{f(\boldsymbol{y},\boldsymbol{y}^*|\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m_1)}, m_1)/f(\boldsymbol{y}^*|\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m_1)}, m_1)}\right) - \psi$$

$$\psi = log\left(\frac{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}}|}{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}}|}\right) - log\left(\frac{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m_1)}}|}{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m_0)}}|}\right) - 2log\left(\frac{f(m_1)}{f(m_0)}\right),$$

where $\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m)}$ is the maximum likelihood estimate using actual and prior data $(\boldsymbol{y},\boldsymbol{y}^*)$ while $\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m_0)}$ is the maximum likelihood estimate using only the prior data $(\boldsymbol{y}^*)$. The determinants involved in the penalty function are given by

$$\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m)}} = -\left[\frac{\partial^2\left\{log\left[f\left(\boldsymbol{y},\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, m\right)\right]\right\}}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\hat{\boldsymbol{\beta}}^{(\boldsymbol{y},\boldsymbol{y}^*)}_{(m)}},$$

and

$$\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}} = -\left[\frac{\partial^2\left\{log\left[f\left(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, m\right)\right]\right\}}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}}.$$

Proof: See appendix, page 207. $\triangleleft$

Alternatively, instead of the above prior, we may use the following 'fractional' prior

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{X}^*_{(m)}, \boldsymbol{y}^*) \propto \left[f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}^*_{(m)}, m)\right]^{1/c_0^2}. \quad (6.40)$$

The parameter c_0^2 controls the weight of information that prior points contribute in the posterior. The actual data (and the likelihood) contribute in the posterior distribution total weight equal to $n/(n+n_0/c_0^2)$ (where n_0 is the size of prior data). When low information is entered it is natural to assume $c_0^2 = n_0$ and therefore the weight of the actual data becomes equal to $n/(n+1)$ while the weight of the prior data is equal to only one data point.

Let us now denote by

$$l(\boldsymbol{y}, \boldsymbol{y}^*, w_1, w_2|m) = \prod_{i=1}^{n}[f(y_i|\boldsymbol{\beta}_{(m)}, m)]^{w_1} \prod_{i=1}^{n_0}[f(y_i^*|\boldsymbol{\beta}_{(m)}, m)]^{w_2}$$

the weighted likelihood of model m with real data $\boldsymbol{y}$, each one having weight equal to w_1, and prior data $\boldsymbol{y}^*$, each one having weight equal to w_2. For example, $l(\boldsymbol{y}, \boldsymbol{y}^*, 0, c_0^{-2}|m)$ is the prior given by (6.41) while $l(\boldsymbol{y}, \boldsymbol{y}^*, 1, 0|m)$ is the usual likelihood. The above considerations lead us to the following proposition.

Proposition 6.16 *The posterior odds of model m_0 against model m_1 using prior of the form*

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, m) \propto [f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}^*_{(m)}, m)]^{1/c_0^2} \tag{6.41}$$

where $\boldsymbol{y}^*$ are imaginary data that express our prior information, can be written in a form equivalent to information criteria given by

$$-2log(PO_{01}) \approx -2log\left(\frac{l(\boldsymbol{y}, \boldsymbol{y}^*, 1, c_0^{-2}|m_0)/l(\boldsymbol{y}, \boldsymbol{y}^*, 0, c_0^{-2}|m_0)}{l(\boldsymbol{y}, \boldsymbol{y}^*, 1, c_0^{-2}|m_1)/l(\boldsymbol{y}, \boldsymbol{y}^*, 0, c_0^{-2}|m_1)}\right) - \psi$$

$$\psi = [d(m_1) - d(m_0)]log(c_0^2) + log\left(\frac{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m_1)}}|}{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)_0}}|}\right) - log\left(\frac{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{*(1/c_0^2)}_{(m_1)}}|}{|\mathcal{I}_{\hat{\boldsymbol{\beta}}^{*(1/c_0^2)}_{(m_0)}}|}\right) - 2log\left(\frac{f(m_1)}{f(m_0)}\right),$$

where $\hat{\boldsymbol{\beta}}^{*(1/c_0^2)}_{(m)}$ is the maximum of the weighted likelihood $l(\boldsymbol{y}, \boldsymbol{y}^*, 1, c_0^{-2}|m)$ and the determinants involved in the penalty are now given by

$$\mathcal{I}_{\hat{\boldsymbol{\beta}}^{*(1/c_0^2)}_{(m)}} = -\left[\frac{\partial^2\left\{log\left[l(\boldsymbol{y}, \boldsymbol{y}^*, 1, c_0^{-2}|m)\right]\right\}}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\hat{\boldsymbol{\beta}}^{*(1/c_0^2)}_{(m)}},$$

$$\mathcal{I}_{\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}} = -\left[\frac{\partial^2\left\{log\left[f\left(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, m\right)\right]\right\}}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}}.$$

Proof: See appendix, page 208. ◁

Without loss of generality we may assume that for every prior density assigned on the model parameters, there exists a prior distribution of type (6.40) for which the product $f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m)}, m) f_N(\breve{\boldsymbol{\beta}}_{(m)}|m)$ can be expressed as a ratio of two maximum weighted likelihoods. The likelihood in the numerator is a measure of fit if information resulting from both prior and data is used while the likelihood of the denominator is the measure of fit resulting if only the prior information is used. The parameter c_0^2 plays similar role as the variance multiplicator c^2 in normal linear models.

Straightforward examples can be given in normal linear and generalised linear models. In normal models a prior of type (6.40) results in

$$f\left(\boldsymbol{\beta}_{(m)}|\sigma^2, \boldsymbol{y}^*, \boldsymbol{X}^*_{(m)}\right) \sim N\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}, c_0^2\left(\boldsymbol{X}^{*T}_{(m)}\boldsymbol{X}^*_{(m)}\right)^{-1}\sigma^2\right).$$

The prior setup (6.10) used by Smith and Kohn (1996) corresponds to an experiment with $n_0 = n$ observations, the same data matrix for both prior and actual data ($\boldsymbol{X}^*_{(m)} = \boldsymbol{X}_{(m)}$), and all the response data equal to zero, $\boldsymbol{y}^* = \mathbf{0}$. The choice $c^2 = n$ in such case is natural since the prior will contribute in the posterior density only by $1/(n+1)$ fraction.

In generalised linear models with likelihood

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, \boldsymbol{X}^*_{(m)}, \phi, m) = exp\left(\sum_{i=1}^{n}\frac{y_i\vartheta_i - b(\vartheta_i)}{a_i(\phi)} + \sum_{i=1}^{n}c(y_i, a_i(\phi))\right)$$

where ϑ_i is a function of the expected value $\mu_i = E[Y_i]$ linked with the parameters of interest, $\boldsymbol{\beta}_{(m)}$, via the link function $g(\eta_i)$ and η is the linear predictor. In this case the prior of type (6.40) is given by

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, \boldsymbol{X}^*_{(m)}, \phi, m) \propto exp\left(\sum_{i=1}^{n_0}\frac{y_i^*\vartheta_i - b(\vartheta_i)}{c_0^2 a_i(\phi)} + c_0^{-2}\sum_{i=1}^{n_0}c(y_i^*, a_i(\phi))\right)$$

which can be well approximated by

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, \boldsymbol{X}^*_{(m)}, \phi, m) \approx N\left(\hat{\boldsymbol{\beta}}^{(\boldsymbol{y}^*)}_{(m)}, c_0^2\left(\boldsymbol{X}^{*T}_{(m)}\boldsymbol{H}^*_{(m)}\boldsymbol{X}^*_{(m)}\right)^{-1}\right),$$

where $\boldsymbol{H}^*_{(m)} = Diag(h_i^*)$ with $h_i^{*-1} = g'(E[Y_i^*])^2 a_i(\phi) v(E[Y_i^*])$; $g(x)$ is the link function used.

If we assume $\boldsymbol{\mu}_{\beta_{(m)}} = \mathbf{0}$ we essentially set $E(Y_i^*) = g^{-1}(0)$. The weights h_i^* involved in matrix $\boldsymbol{H}^*_{(m)}$ are now given by

$$h_i^* = \{[g'(g^{-1}(0))]^2 a_i(\phi) v(g^{-1}(0))\}^{-1}.$$

Usually $a_i(\phi) = \phi/w_i$; where w_i are weights for each observation (equal to one when ungrouped data are used). Here we assume that our prior data are ungrouped and hence $w_i = 1$ resulting in $h_i^* = \{\phi[g'(g^{-1}(0))]^2 v(g^{-1}(0))\}^{-1}$ which is constant over all prior data.

Therefore the prior distribution of type

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, \boldsymbol{X}^*_{(m)}, \phi, m) = N\left(\boldsymbol{0}, c^2\left(\boldsymbol{X}^{*T}_{(m)}\boldsymbol{X}^*_{(m)}\right)^{-1}\phi\right)$$

corresponds to n_0 prior points, each one weighted by $1/c_0^2$, data matrix $\boldsymbol{X}^*_{(m)}$, $\boldsymbol{y}^* = \boldsymbol{0}$ and $c_0^2 = c^2\{[g'(g^{-1}(0))]^2 v(g^{-1}(0))\}^{-1}$. When the design matrix of the prior data is set equal to $\boldsymbol{X}^*_{(m)} = \boldsymbol{X}_{(m)}$ then $n_0 = n$ and hence plausible choice for $c^2 = n[g'(g^{-1}(0))]^2 v(g^{-1}(0))$ which results in $n/(n+1)$ contribution of the data and $1/(n+1)$ of the prior information.

6.6.2 Prior Distributions via Penalty Determination in Generalised Linear Models

In the section which follows we specify the prior distributions via prior determination of the desired penalty imposed to the prior and posterior weighted maximum likelihood ratios presented in proposition 6.16.

Proposition 6.17 *The posterior odds of model m_0 against model m_1 with prior distribution*

$$f(\boldsymbol{\beta}_{(m)}|m) \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, c^2\boldsymbol{V}_{(m)}\right)$$

for $m \in \{m_0, m_1\}$ can be written in a form equivalent to information criteria given by

$$-2log(PO_{01}) \approx -2log\left(\frac{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_0)}, m_0) f_N(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)}{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_1)}, m_1) f_N(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)}\right) - \psi$$

with penalty function

$$\psi = \{d(m_1) - d(m_0)\}log(c^2) + log\left(\frac{|\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{V}_{(m_0)}|}\right) + log\left(\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}_{(m_0)}|}\right) - 2log\left(\frac{f(m_1)}{f(m_0)}\right),$$

where $\boldsymbol{H}_{(m)} = Diag(h_i)$, $h_i = \{g'(E[Y_i])^2 a_i(\phi) v(E[Y_i])\}^{-1}$.

Proof: See appendix, page 208. ◁

The above penalty function is similar to the penalty of the general case of normal models given by (6.9).

Corollary 6.17.1 *The posterior odds of model m_0 against model m_1 with prior distribution*

$$\boldsymbol{\beta}_{(m)} \sim N\left(\mathbf{0}, c^2\Big(\boldsymbol{X}_{(m)}^T \boldsymbol{H}_{(m)} \boldsymbol{X}_{(m)}\Big)^{-1}\right)$$

for $m \in \{m_0, m_1\}$ can be written in a form equivalent to information criteria given by

$$-2log(PO_{01}) \approx -2log\left(\frac{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_0)}, m_0) f_N(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)}{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_1)}, m_1) f_N(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)}\right) - \psi$$

$$\psi = \{d(m_1) - d(m_0)\} log(c^2+1) - 2log\left(\frac{f(m_1)}{f(m_0)}\right),$$

where $\boldsymbol{H}_{(m)} = Diag(h_i)$, $h_i = \{g'(E[Y_i])^2 a_i(\phi) v(E[Y_i])\}^{-1}$.

Proof: See appendix, page 208. $\triangleleft$

For large c^2, the posterior mode becomes approximately equal to the maximum likelihood estimate and the ratio $f_N(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)/f_N(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)$ will be approximately equal to one. In such cases the $-2log(PO_{01})$ will be equivalent to information criteria with the above penalty functions. If we apply the prior specification method proposed in Section 6.4.1 then we have the following proposition.

Proposition 6.18 *Consider two generalised linear models m_0 and m_1 with prior distributions*

$$f(\boldsymbol{\beta}_{(m)}|m) \sim N\left(\boldsymbol{\mu}_{\beta_{(m)}}, c^2 \boldsymbol{V}_{(m)}\right)$$

for both $m \in \{m_0, m_1\}$ and prior model probabilities

$$f(m) \propto \left(\frac{c^2}{\kappa}\right)^{d(m)/2} |\boldsymbol{V}_{(m)}|^{1/2} |\boldsymbol{X}_{(m)}^T \boldsymbol{H}_{(m)} \boldsymbol{X}_{(m)} + c^{-2} \boldsymbol{V}_{(m)}|^{1/2}.$$

Then the posterior odds of model m_0 against model m_1 can be written in a form equivalent to information criteria given by

$$-2log(PO_{01}) \approx -2log\left(\frac{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_0)}, m_0) f_N(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)}{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_1)}, m_1) f_N(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)}\right) - \psi$$

with penalty function

$$\psi = \{d(m_1) - d(m_0)\} log(\kappa).$$

Proof: See appendix, page 208. $\triangleleft$

Similarly to the above proposition we have the following corollary for the simpler case.

Corollary 6.18.1 *Consider two generalised linear models m_0 and m_1 with prior distributions*

$$\boldsymbol{\beta}_{(m)} \sim N\left(\mathbf{0}, c^2\Big(\boldsymbol{X}_{(m)}^T\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m)}\Big)^{-1}\right)$$

for both $m \in \{m_0, m_1\}$ with prior model probabilities

$$f(m) \propto \left(\frac{c^2+1}{\kappa}\right)^{d(m)/2}.$$

Then the posterior odds of model m_0 against model m_1 can be written in a form equivalent to information criteria given by

$$-2log(PO_{01}) \approx -2log\left(\frac{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_0)}, m_0)f_N(\breve{\boldsymbol{\beta}}_{(m_0)}|m_0)}{f(\boldsymbol{y}|\breve{\boldsymbol{\beta}}_{(m_1)}, m_1)f_N(\breve{\boldsymbol{\beta}}_{(m_1)}|m_1)}\right) - \psi$$

with penalty function

$$\psi = \{d(m_1) - d(m_0)\}log(\kappa).$$

Proof: See appendix, page 209. $\triangleleft$

The prior odds at zero are similar to the normal linear model. The posterior odds at zero can be calculated using χ^2 distribution.

6.7 Bayes Factor's Variants and Information Criteria

Here we briefly review the association of the three most popular variants of Bayes factor (posterior, intrinsic and fractional) with the information criteria and we further investigate the behaviour of the SSVS Bayes factor under certain conditions.

6.7.1 Posterior, Fractional and Intrinsic Bayes Factors.

The need to use non-informative priors in model selection led to the definition of three new types of Bayes factors: the posterior, the fractional and the intrinsic Bayes factors by Aitkin (1991), O'Hagan (1995) and Berger and Pericchi (1996a, 1996b), respectively. Here we briefly review the association of Bayes factor with information criteria.

According to O'Hagan (1995) fractional Bayes factor can be written as

$$-2log(FBF_{b,01}) = -2(1-b)log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_0}, m_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_1}, m_1)}\right) - \{\ d(m_1) - d_(m_0)\ \}log(1/b) \qquad (6.42)$$

where $b<1$ is the fractional parameter. It is obvious that the log-likelihood ratio test is penalised by the fractional parameter. Moreover, instead of the full log-likelihood ratio we use a fraction of it depending on parameter b. The posterior Bayes factor, introduced by Aitkin (1991), is even more closely related to information criteria since it is given by

$$-2log(PBF_{01}) = -2log\left(\frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_0}, m_0)}{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{m_0}, m_0)}\right) - \{\ d(m_1) - d(m_2)\ \}log(2), \tag{6.43}$$

that is $F = log(2)$. This penalty is quite small compared to AIC in which $F=2$ and BIC in which $F = log(n)$ and therefore posterior Bayes factor supports more complicated models. Note that BIC gives the same penalty only for samples with only two observations (minimal required sample for estimating variance). Finally, it is clear that there is a prior for which the corresponding usual Bayes factor is exactly the same as the posterior Bayes factor. If we use the prior (6.10) with $c^2=1$ then the penalizing part is the same as in posterior Bayes factor but the ratio SS_{m_1}/SS_{m_0} will no longer be equal to likelihood ratio RSS_{m_1}/RSS_{m_0}. On the other hand, if we adopt the setup of Section 6.4.1 with large c^2 and $\kappa=2$ then we have a Bayes factor which is the same with the posterior Bayes factor without using any information from the data $\boldsymbol{y}$.

Berger and Pericchi (1996a) introduced the intrinsic Bayes factor. Generally, intrinsic Bayes factor cannot be written in the general form of information criteria given by (6.5). Intrinsic Bayes factor using Jeffreys prior is the only one from the three different improper priors used in normal linear models by Berger and Pericchi (1996b) which results to an intrinsic Bayes factor closely related to information criteria as defined above. The resulted criterion has the form of (6.5) with penalty

$$\psi = log\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}|} + 2log\left[L(n_0)^{-1}\sum_{l\in\mathcal{L}(n_0)}\left\{\left(\frac{|\boldsymbol{X}^T_{(m_0)}(l)\boldsymbol{X}_{(m_0)}(l)|}{|\boldsymbol{X}^T_{(m_1)}(l)\boldsymbol{X}_{(m_1)}(l)|}\right)^{1/2}\left(\frac{RSS_{m_0}(l)}{RSS_{m_1}(l)}\right)^{n_0/2}\right\}\right]$$

for arithmetic intrinsic Bayes factor and

$$\psi = log\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}|} + \frac{1}{L(n_0)}\sum_{l\in\mathcal{L}(n_0)} log\frac{|\boldsymbol{X}^T_{(m_0)}(l)\boldsymbol{X}_{(m_0)}(l)|}{|\boldsymbol{X}^T_{(m_1)}(l)\boldsymbol{X}_{(m_1)}(l)|} + \frac{n_0}{L(n_0)}\sum_{l\in\mathcal{L}(n_0)} log\frac{RSS_{m_0}(l)}{RSS_{m_1}(l)}$$

for the geometric intrinsic Bayes factor; where n_0 denotes the size of the training sample, $RSS_m(l)$ and $\boldsymbol{X}_{(m)}(l)$ are the residual sum of squares and the design matrix of model m using training sample $\boldsymbol{y}(l)$ respectively, $\mathcal{L}(n_0)$ and $L(n_0)$ are the set and the number of all possible training samples of size n_0 respectively.

6.7.2 The SSVS Bayes Factor.

SSVS was introduced by George and McCulloch (1993); for details see Section 3.4.1. The basic idea of SSVS is the use of prior distributions of the form

$$\boldsymbol{\beta}_j|\gamma_j \sim \gamma_j N(0, \boldsymbol{\Sigma}_j) + (1-\gamma_j)N(0, k_j^{-2}\boldsymbol{\Sigma}_j)$$

where γ_j is a binary variable indicator and k_j^2 is defined according to our prior beliefs for significant and non-significant limits; see George and McCulloch (1993) semi-automatic selection. Usually k_j is substituted by a common k for all terms. Generally, SSVS gives different results than traditional model selection methods since it uses different likelihood. When $k_j^2 \to \infty$, $k_j^{-2}\boldsymbol{\Sigma}_j \to \mathbf{0}_{d_j}$ and hence the prior becomes

$$\boldsymbol{\beta}_j|\gamma_j \sim \gamma_j N(\mathbf{0}, \boldsymbol{\Sigma}_j) + (1-\gamma_j)I(\mathbf{0}_{d_j})$$

where $I(\mathbf{0}_{d_j})$ is a mass prior at zero and therefore the posterior becomes the same as traditional model selection methods; for details see George and McCulloch (1997). For this reason the posterior odds estimated by SSVS are approximately equal (for large k) to the common posterior odds. We denote this posterior odds as PO_{01}^{SSVS} and the corresponding Bayes factor as B_{01}^{SSVS}. Generally we have that

$$PO_{01}^{SSVS} \to PO_{01}, \quad when \;\; k^2 \to \infty.$$

In this section we will use the general prior setup

$$f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\gamma}) \sim N\left(\mathbf{0}, c^2\boldsymbol{V}_{(m)}^{SSVS}\right), \quad \boldsymbol{V}_{(m)}^{SSVS} = D(k^{\gamma_j-1}I_{d_j})\boldsymbol{R}D(k^{\gamma_j-1}I_{d_j}) \text{ and } f(\sigma^2) \propto \sigma^{-2} \tag{6.44}$$

where d is the dimension of the full model and $D(k^{\gamma_j-1}I_{d_j})$ is a $d \times d$ block diagonal matrix with diagonal elements equal to the identity matrix of dimension d_j multiplied by k^{-1} if the corresponding j term is excluded from the model This generalised prior setup was also proposed by George and McCulloch (1993). Interest lies in special families of prior distributions which correspond to the ones examined in the case of simple model comparison; that is $\boldsymbol{R}^{-1} = \boldsymbol{X}^T\boldsymbol{X}$ for the Smith and Kohn (1996) prior setup and $\boldsymbol{R}^{-1} = D(\boldsymbol{X}_j\boldsymbol{X}_j)$ for the independent prior setup.

6.7.2.1 The General Model Comparison

The calculation of posterior odds are straightforward following the calculation of Section 6.2.1 resulting to

$$-2log(PO_{01}^{SSVS}) = nlog\left(\frac{SS_{m_0}^{SSVS}}{SS_{m_1}^{SSVS}}\right) - \psi, \tag{6.45}$$

$$\psi = log\left(\frac{|\boldsymbol{V}_{(m_1)}^{SSVS}|}{|\boldsymbol{V}_{(m_0)}^{SSVS}|}\right) + log\left(\frac{|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}_{(m_1)}^{SSVS}]^{-1}|}{|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}_{(m_0)}^{SSVS}]^{-1}|}\right) - 2log\left(\frac{f(m_1)}{f(m_0)}\right). \tag{6.46}$$

where SS_m^{SSVS} is the SSVS based posterior sum of squares given by

$$SS_m^{SSVS} = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}_{(m)}^{SSVS}]^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

We are going to examine the association of the above posterior odds and the usual posterior odds when k^2 is large. Without loss of generality, for every model m we can write

$$\boldsymbol{X} = [\boldsymbol{X}_{(m)}\boldsymbol{X}_{(\backslash m)}], \quad \boldsymbol{\beta}^T = [\boldsymbol{\beta}_{(m)}^T\boldsymbol{\beta}_{(\backslash m)}^T], \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{(m)} & \boldsymbol{R}_{(m,\backslash m)} \\ \boldsymbol{R}_{(m,\backslash m)}^T & \boldsymbol{R}_{(\backslash m)} \end{bmatrix}$$

where $\boldsymbol{X}_{(\backslash m)}$ and $\boldsymbol{\beta}_{(\backslash m)}$ refers to the components of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ that are associated with terms excluded from model m. The matrix $\boldsymbol{R}$ is partitioned to the matrices $\boldsymbol{R}_{(m)}$ that corresponds to covariances between terms included in model m, $\boldsymbol{R}_{(\backslash m)}$ that corresponds to covariances between terms excluded from model m and $\boldsymbol{R}_{(m,\backslash m)}$ that corresponds to covariances between each term included in model m with each term excluded from model m.

Proposition 6.19 *Consider two normal linear models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n\right)$$

and prior setup

$$f(\boldsymbol{\beta}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2D(k^{\gamma_j-1}I_{d_j})\boldsymbol{R}D(k^{\gamma_j-1}I_{d_j})\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$. Then

$$\lim_{k^2/c^2\to\infty}(PO_{01}^{SSVS}) = PO_{01}$$

where PO_{01}^{SSVS} and PO_{01} are the posterior odds for SSVS and usual model selection with prior matrix $\boldsymbol{V}_{(m)}$ given by

$$\boldsymbol{V}_{(m)} = \boldsymbol{R}_{(m)}.$$

Proof: See appendix, page 209. $\triangleleft$

The above proposition clearly states that using such prior for large k^2 is equivalent to setting a prior matrix $\boldsymbol{V}_{(m)} = \boldsymbol{R}_{(m)}$ for each model in the usual model selection.

6.7.2.2 Lindley-Bartlett's Paradox and SSVS

In this section we briefly present the behaviour of the Bayes factor when the sample size or the prior parameter c^2 tend to infinity, for fixed k^2.

Proposition 6.20 *Consider two normal linear models m_0 and m_1 given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right)$$

and prior setup

$$f(\boldsymbol{\beta}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 D(k^{\gamma_j-1} I_{d_j})\boldsymbol{R}D(k^{\gamma_j-1} I_{d_j})\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$. Then, under mild regularity conditions,

$$\lim_{n\to\infty}\left(-2log(B_{01}^{SSVS})\right) = \{d(m_1) - d(m_0)\}log(k^2)$$

where B_{01}^{SSVS} is the Bayes factors for SSVS.

Proof: See appendix, page 212. $\triangleleft$

Proposition 6.21 *Consider two normal linear models m_0 and m_1 which are given by*

$$\boldsymbol{y} \sim N\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right)$$

and prior setup given by

$$f(\boldsymbol{\beta}|\sigma^2, m) \sim N\left(\boldsymbol{0}, c^2 D(k^{\gamma_j-1} I_{d_j})\boldsymbol{R}D(k^{\gamma_j-1} I_{d_j})\sigma^2\right) \text{ and } f(\sigma^2) \propto \sigma^{-2}$$

for both $m \in \{m_0, m_1\}$. Then

$$\lim_{c^2\to\infty}\left(-2log(B_{01}^{SSVS})\right) = \{d(m_1) - d(m_0)\}log(k^2)$$

where B_{01}^{SSVS} is the Bayes factors for SSVS.

Proof: See appendix, page 212. $\triangleleft$

A paradox similar to Lindley-Bartlett paradox also occurs for SSVS Bayes factor. The SSVS Bayes factor is bounded and this bound depends on the magnitude of k^2. Therefore, the SSVS based Bayes factor does not avoid the Lindley-Bartlett paradox which still appears in a slightly different form.

6.8 Discussion

In this chapter we have presented some problems and possible solutions regarding the use of 'non-informative' priors in Bayesian model selection. Some specific model selection setups in linear and generalised linear models have been presented. Also, the connection of posterior odds and information criteria was reported. This connection was used to specify the prior probabilities in order to achieve a desired penalty and remove the prior variance effect. The notion of conditional prior odds was also discussed and implemented in simple normal linear model examples. Bayes factor's variants were also discussed including SSVS based Bayes factor and its limiting behaviour.

We argue that the uniform prior on model space could be avoided. The combination of the prior odds and the prior variance of model parameters may not reflect our real prior beliefs for models and may support more complicated or simpler models than the ones a-priori desired. Instead, a joint specification of both prior odds and prior variance is advocated. Prior probabilities can be determined as a function of dimensionality and a new parameter which controls the imposed penalty. Finally, implementation in generalised linear models via the use of Laplace approximations is discussed in detail.

6.9 Appendix: Proofs

Proof of Proposition 6.1. From equations (6.3) we have that

$$\tilde{\boldsymbol{\Sigma}}_{(m)}^{-1} = \boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{V}_{(m)}^{-1}$$

which for large c^2 becomes approximately equal to

$$\tilde{\boldsymbol{\Sigma}}_{(m)}^{-1} \approx \boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}.$$

Similarly

$$\tilde{\boldsymbol{\beta}}_{(m)} \approx \left(\boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}\right)^{-1} \boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)} \hat{\boldsymbol{\beta}}_{(m)} = \hat{\boldsymbol{\beta}}_{(m)}.$$

Substituting the above two approximations in posterior sum of squares given by (6.2) we have

$$SS_m \approx \boldsymbol{y}^T \boldsymbol{y} - \hat{\boldsymbol{\beta}}_{(m)}^T \boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)} \tilde{\boldsymbol{\beta}}_{(m)} = RSS_m.$$

Finally, if we consider the above approximation and the posterior odds (6.1) we have

$$-2log(PO_{01}) \approx nlog\left(\frac{RSS_{m_0}}{RSS_{m1}}\right) - \{\ d(m_1) - d(m_0)\ \}log(c^2) +$$

$$+log\left(\frac{|\boldsymbol{V}_{(m_0)}|}{|\boldsymbol{V}_{(m_1)}|}\right) + log\left(\frac{|\boldsymbol{X}_{(m_0)}^T \boldsymbol{X}_{(m_0)}|}{|\boldsymbol{X}_{(m_1)}^T \boldsymbol{X}_{(m_1)}|}\right) + 2log\left(\frac{f(m_1)}{f(m_0)}\right).$$

which is equivalent to an information criterion (6.6) [or (6.5)] with penalty function given by (6.8). $\triangleleft$

Proof of Proposition 6.2. The prior setup of proposition 6.2 leads to posterior odds given by

$$PO_{01} = \frac{f(m_0)}{f(m_1)} \frac{f(\boldsymbol{y}|m_0)}{f(\boldsymbol{y}|m_1)} = (c^2+1)^{\{d(m_1)-d(m_0)\}/2} \left(\frac{SS_{m_0}}{SS_{m_1}}\right)^{-n/2}$$

resulting in

$$-2log(PO_{01}) = nlog\left(\frac{SS_{m_0}}{SS_{m_1}}\right) - \{d(m_1) - d(m_0)\}log(c^2+1) - 2log\left(\frac{f(m_1)}{f(m_0)}\right).$$

Under the above prior setup

$$\tilde{\boldsymbol{\beta}}_{(m)} = \frac{c^2}{c^2+1}\left(\boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}\right)^{-1} \boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)} \hat{\boldsymbol{\beta}}_{(m)} = \frac{c^2}{c^2+1}\hat{\boldsymbol{\beta}}_{(m)},$$

$$\tilde{\boldsymbol{\Sigma}}_{(m)} = \frac{c^2}{c^2+1}\left(\boldsymbol{X}_{(m)}^T \boldsymbol{X}_{(m)}\right)^{-1},$$

and therefore the posterior sum of squares are simplified to

$$SS_m = \boldsymbol{y}^T\boldsymbol{y} - \frac{c^2}{c^2+1}\hat{\boldsymbol{\beta}}^T_{(m)}\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\right)^{-1}\hat{\boldsymbol{\beta}}_{(m)}$$

which, for large c^2, becomes equal to the residual sum of squares:

$$SS_m \approx RSS_m \text{ for large } c^2$$

and therefore $IC_{01} \approx -2log(PO_{01})$ with penalty function ψ given by (6.11). $\triangleleft$

Proof of Proposition 6.3. In this case $\boldsymbol{V}_{(m)} = D^{-1}_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)$ and therefore its determinant is given by the product of the determinants $|\boldsymbol{X}^T_j\boldsymbol{X}_j|^{-1}$ for all terms included in model m. Therefore

$$|\boldsymbol{V}_{(m)}| = |D^{-1}_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)| = |D_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)|^{-1} = \prod_{j'=1}^{p}|\boldsymbol{X}^T_{j'}\boldsymbol{X}_{j'}|^{-\gamma_{j',m}}.$$

Since the two models differ by only j term we have that

$$\frac{|\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{V}_{(m_0)}|} = \prod_{j'=1}^{p}|\boldsymbol{X}^T_{j'}\boldsymbol{X}_{j'}|^{-\gamma_{j',m_1}+\gamma_{j',m_0}} = |\boldsymbol{X}^T_j\boldsymbol{X}_j|^{-1}.$$

The matrix $\tilde{\boldsymbol{\Sigma}}_{(m_1)}$ is given by

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}^{-1}_{(m_1)} &= \boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}D_{(m_1)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\\
&= [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j]^T[\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] + c^{-2}D_{(m_1)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)\\
&= \begin{bmatrix} \boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}^T_j\boldsymbol{X}_j) & \boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j \\ \boldsymbol{X}^T_j\boldsymbol{X}_{(m_0)} & \frac{c^2+1}{c^2}\boldsymbol{X}^T_j\boldsymbol{X}_j \end{bmatrix}\\
&= \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}^{-1}_{(m_0)} & \boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j \\ \boldsymbol{X}^T_j\boldsymbol{X}_{(m_0)} & \frac{c^2+1}{c^2}\boldsymbol{X}^T_j\boldsymbol{X}_j \end{bmatrix}.
\end{aligned}$$

From the properties of partitioned matrices we have that

$$|\tilde{\boldsymbol{\Sigma}}^{-1}_{(m_1)}| = |\tilde{\boldsymbol{\Sigma}}^{-1}_{(m_0)}|\left|\frac{c^2+1}{c^2}\boldsymbol{X}^T_j\boldsymbol{X}_j - \boldsymbol{X}^T_j\boldsymbol{X}_{(m_0)}\tilde{\boldsymbol{\Sigma}}_{(m_0)}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j\right|,$$

resulting in

$$\begin{aligned}
\frac{|\tilde{\boldsymbol{\Sigma}}^{-1}_{(m_1)}|}{|\tilde{\boldsymbol{\Sigma}}^{-1}_{(m_0)}|} &= \frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}D_{(m_1)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)|}\\
&= \left(\frac{c^2+1}{c^2}\right)^{d_j}\left|\boldsymbol{X}^T_j\boldsymbol{X}_j - \frac{c^2}{c^2+1}\boldsymbol{X}^T_j\boldsymbol{X}_{(m_0)}\tilde{\boldsymbol{\Sigma}}_{(m_0)}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j\right|\\
&= \left(\frac{c^2+1}{c^2}\right)^{d_j}\left|\boldsymbol{X}^T_j\left(\boldsymbol{I}_n - \frac{c^2}{c^2+1}\boldsymbol{X}_{(m_0)}\tilde{\boldsymbol{\Sigma}}_{(m_0)}\boldsymbol{X}^T_{(m_0)}\right)\boldsymbol{X}_j\right|\\
&= \left(\frac{c^2+1}{c^2}\right)^{d_j}\left|\boldsymbol{X}^T_j\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j\right|.
\end{aligned}$$

Using the above equations in the penalty function (6.8) we have the result of proposition 6.3. $\triangleleft$

Proof of Corollary 6.3.1. The proof is immediate from proposition 6.3 since $\boldsymbol{X}_j^T \boldsymbol{X}_{(m_0)} = \boldsymbol{0}$ due to the assumed orthogonality. Therefore $\boldsymbol{X}_j^T \boldsymbol{\Delta}_{(m_0)} \boldsymbol{X}_j = \boldsymbol{X}_j^T \boldsymbol{X}_j$ resulting to the penalty of corollary 6.3.1 $\triangleleft$

Proof of Proposition 6.4. In regression R^2 coefficient is given by

$$R^2 = 1 - \frac{RSS_m}{(n-1)s_y^2}.$$

If we use as response the variable X_j then we have

$$R^2_{m_0,x_j} = 1 - \frac{RSS_{m_0,x_j}}{(n-1)s_j^2},$$

where RSS_{m_0,x_j} is the residual sum of squares of a regression with response X_j and explanatory variables all terms included in model m_0. The scalar $\boldsymbol{X}_j^T \boldsymbol{\Delta}_{(m_0)} \boldsymbol{X}_j$ for large c^2 is approximately equal to the posterior sum of squares if X_j is used as response. From equation (6.4) the posterior sum of squares SS_{m_0,x_j} with response $\boldsymbol{X}_j$, design matrix $\boldsymbol{X}_{(m_0)}$ and prior distribution (6.12) is given by

$$\begin{aligned} SS_{m_0,x_j} &= RSS_{m_0,x_j} \\ &\quad + \hat{\boldsymbol{\beta}}^T_{(m_0,x_j)} \left[\left(\boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_{(m_0)} \right)^{-1} + c^2 D^{-1}_{(m_0)} (\boldsymbol{X}_j^T \boldsymbol{X}_j) \right]^{-1} \hat{\boldsymbol{\beta}}_{(m_0,x_j)} \\ &= (n-1)s_j^2(1 - R^2_{m_0,x_j}) \\ &\quad + \hat{\boldsymbol{\beta}}^T_{(m_0,x_j)} \left[\left(\boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_{(m_0)} \right)^{-1} + c^2 D^{-1}_{(m_0)} (\boldsymbol{X}_j^T \boldsymbol{X}_j) \right]^{-1} \hat{\boldsymbol{\beta}}_{(m_0,x_j)} \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_{(m_0,x_j)}$ is the vector of maximum likelihood estimates of the coefficients of the regression model with response X_j and covariates X_ν for all $\nu \in \mathcal{V}(m_0)$ while $R^2_{m_0,x_j}$ is the R^2 measure resulted from a regression model with response the additional variable X_j and covariates all X_ν for $\nu \in \mathcal{V}(m_0)$. Therefore, we have that

$$\begin{aligned} \boldsymbol{X}_j^T \boldsymbol{\Delta}_{(m_0)} \boldsymbol{X}_j &= \boldsymbol{X}_j^T \boldsymbol{X}_j - \frac{c^2}{c^2+1} \boldsymbol{X}_j^T \boldsymbol{X}_{(m_0)} \tilde{\boldsymbol{\Sigma}}_{(m_0)} \boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_j \\ &= \boldsymbol{X}_j^T \boldsymbol{X}_j - \boldsymbol{X}_j^T \boldsymbol{X}_{(m_0)} \tilde{\boldsymbol{\Sigma}}_{(m)} \boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_j \\ &\quad + \frac{1}{c^2+1} \boldsymbol{X}_j^T \boldsymbol{X}_{(m_0)} \tilde{\boldsymbol{\Sigma}}_{(m)} \boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_j \\ &= SS_{m_0,x_j} + \frac{1}{c^2+1} \boldsymbol{X}_j^T \boldsymbol{X}_{(m_0)} \tilde{\boldsymbol{\Sigma}}_{(m)} \boldsymbol{X}^T_{(m_0)} \boldsymbol{X}_j \end{aligned}$$

$$
\begin{aligned}
&= (n-1)s_j^2(1-R^2_{m_0,x_j}) \\
&\quad +\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)}\left[\left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\right)^{-1}+c^2D^{-1}_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right]^{-1}\hat{\boldsymbol{\beta}}_{(m_0,x_j)} \\
&\quad +\frac{1}{c^2+1}\boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\tilde{\boldsymbol{\Sigma}}_{(m)}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j
\end{aligned}
$$

where $\tilde{\boldsymbol{\Sigma}}_{(m)}$ in the above equations is given by

$$
\tilde{\boldsymbol{\Sigma}}_{(m)}=\left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}+c^{-2}D_{(m)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right)^{-1}.
$$

In the collinear case $R^2_{m_0,x_j}=1$ since $X_j=\boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0,x_j)}$. Therefore, we have that

$$
\begin{aligned}
\boldsymbol{X}_j^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j=0 \quad &+\quad \hat{\boldsymbol{\beta}}^T_{(m_0,x_j)}\left[\left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\right)^{-1}+c^2D^{-1}_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right]^{-1}\hat{\boldsymbol{\beta}}_{(m_0,x_j)} \\
&+\quad \frac{1}{c^2+1}\boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\tilde{\boldsymbol{\Sigma}}_{(m_0)}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j.
\end{aligned}
$$

If we multiply both sides of the equation by the scalar (c^2+1) we have

$$
\begin{aligned}
(c^2+1)\boldsymbol{X}_j^T\boldsymbol{\Delta}_{(m_0)}\boldsymbol{X}_j &= \frac{c^2+1}{c^2}\hat{\beta}^T_{(m_0,x_j)}\left(c^{-2}\left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\right)^{-1}+\frac{c^2}{c^2+1}D^{-1}_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right)^{-1}\hat{\boldsymbol{\beta}}_{(m_0,x_j)} \\
&\quad +\boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}+c^{-2}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right)^{-1}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j \\
&\approx^1 \hat{\beta}^T_{(m_0,x_j)}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\hat{\beta}_{(m_0,x_j)}+\boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\right)^{-1}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_j \\
&= \hat{\beta}^T_{(m_0,x_j)}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\hat{\beta}_{(m_0,x_j)}+\boldsymbol{X}_j^T\boldsymbol{X}_{(m_0)}\hat{\beta}_{(m_0,x_j)} \\
&=^2 \hat{\beta}^T_{(m_0,x_j)}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\hat{\beta}_{(m_0,x_j)}+\boldsymbol{X}_j^T\boldsymbol{X}_j.
\end{aligned}
$$

If we substitute the above result to the penalty (6.13), we have

$$
\psi = log[R^*(m_0,X_j)+1]+2log\left(\frac{f(m_0)}{f(m_1)}\right)
$$

with

$$
R^*(m_0,x_j)=\frac{\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\hat{\boldsymbol{\beta}}_{(m_0,x_j)}}{\boldsymbol{X}_j^T\boldsymbol{X}_j}.
$$

Furthermore, the second expression of $R^*(m_0,x_j)$ is given if we substitute $\boldsymbol{X}_j$ with $\boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0,x_j)}$ due to the assumed collinearity.

The third expression of $R^*(m_0,x_j)$ is obtained if we analyse the quadratic forms of (6.16). Therefore, we may write

$$
\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\hat{\boldsymbol{\beta}}_{(m_0,x_j)} \quad =^3 \quad \sum_{\nu\in\mathcal{V}(m_0)}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]^T_\nu\boldsymbol{X}^T_\nu\boldsymbol{X}_\nu[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_\nu=\alpha_2^*
$$

[1] For large c^2.

[2] Due to collinearity $X_j=\boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0,x_j)}$.

[3] Due to prior distribution (6.12)

and

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^T_{(m_0,x_j)}\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0,x_j)} &= \sum_{\nu\in\mathcal{V}(m_0)}\sum_{\nu'\in\mathcal{V}(m_0)}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]^T_\nu \boldsymbol{X}^T_\nu \boldsymbol{X}_{\nu'}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_{\nu'} \\
&= \sum_{\nu\in\mathcal{V}(m_0)}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]^T_\nu \boldsymbol{X}^T_\nu \boldsymbol{X}_{\nu}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_{\nu} \\
&\quad + \sum_{\nu\in\mathcal{V}(m_0)}\sum_{\nu'\in\mathcal{V}(m_0)\setminus\{\nu\}}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]^T_\nu \boldsymbol{X}^T_\nu \boldsymbol{X}_{\nu'}[\hat{\boldsymbol{\beta}}_{(m_0,x_j)}]_{\nu'} \\
&= \alpha_1^* + \alpha_2^*
\end{aligned}$$

resulting in $R^*(m_0, x_j) = 1 - \alpha_1^*/(\alpha_1^* + \alpha_2^*)$.

Moreover, in the case of collinearity the log-likelihood is maximized if we maximize the quadratic form

$$l = (\boldsymbol{y} - \boldsymbol{X}_{(m_1)}\boldsymbol{\beta}_{(m_1)})^T(\boldsymbol{y} - \boldsymbol{X}_{(m_1)}\boldsymbol{\beta}_{(m_1)}).$$

Since we assume that m_1 contains all terms of model m_0 and the additional term X_j then, without loss of generality, we can write

$$\boldsymbol{X}_{(m_1)} = [\boldsymbol{X}_{(m_0)}, \boldsymbol{X}_j] \text{ and } \boldsymbol{\beta}^T_{(m_1)} = [\boldsymbol{\beta}^*_{(0)}\boldsymbol{\beta}^*_j]$$

where $\boldsymbol{\beta}^*_{(0)}$ is the parameter vector of model m_1 corresponding to terms also included in model m_0, and $\boldsymbol{\beta}^*_j$ is the parameter vector of the additional term j of model m_1. Now we can write

$$l = (\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}^*_{(0)} - \boldsymbol{X}_j\boldsymbol{\beta}^*_j)^T(\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\boldsymbol{\beta}^*_{(0)} - \boldsymbol{X}_j\boldsymbol{\beta}^*_j).$$

The maximum likelihood estimate can now be written as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^*_{(0)} &= \left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\right)^{-1}\boldsymbol{X}_{(m_0)}\boldsymbol{y} - \left(\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}\right)^{-1}\boldsymbol{X}_{(m_0)}\boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j \\
&= \hat{\boldsymbol{\beta}}_{(m_0)} + \hat{\boldsymbol{\beta}}_{(m_0,x_j)}\hat{\boldsymbol{\beta}}^*_j.
\end{aligned}$$

The residual sum of squares are given if we substitute the above quantity on the log-likelihood l and therefore we have

$$RSS_{m_1} = (\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}^*_{(0)} - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j)^T(\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}^*_{(0)} - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j).$$

From the above maximum likelihood solution we have that

$$(\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}^*_{(0)} - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j) = (\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0)} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0,x_j)}\hat{\boldsymbol{\beta}}^*_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j).$$

Finally, since X_j is collinear to the terms of model m_0 we have that $\boldsymbol{X}_j = \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{m_0,x_j}$ and therefore

$$
\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}^*_{(0)} - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j) &= (\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0)} - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}^*_j) \\
&= (\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0)})
\end{aligned}
$$

resulting in

$$
RSS_{m_1} = (\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0)})^T(\boldsymbol{y} - \boldsymbol{X}_{(m_0)}\hat{\boldsymbol{\beta}}_{(m_0)}) = RSS_{m_0}.
$$

For large c^2 we have that $SS_m \approx RSS_m$, $SS_{m_1} \approx SS_{m_0}$ and hence $2log(P0_{01}) = \psi$. $\triangleleft$

Proof of Corollary 6.4.1. The proof is immediate from proposition 6.4 since α_1^* will be equal to zero if $\boldsymbol{X}_{(m_0)}$ is orthogonal. $\triangleleft$

Proof of Proposition 6.5. The conditional prior odds are given by

$$
\begin{aligned}
CPONZ_\xi &= \frac{f(m_1 \mid \boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2, \sigma^2)}{f(m_0 \mid \boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2, \sigma^2)} \\
&= \frac{f(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2|\sigma^2, m_1)f(\sigma^2|m_1)f(m_1)}{f(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2|\sigma^2, m_0)f(\sigma^2|m_0)f(m_0)} \\
&=^1 f(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2|\sigma^2, m_1)\frac{f(\sigma^2|m_1)f(m_1)}{f(\sigma^2|m_0)f(m_0)} \\
&= f(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2|\sigma^2, \gamma_j = 1)\frac{f(\sigma^2|m_1)f(m_1)}{f(\sigma^2|m_0)f(m_0)}. \triangleleft
\end{aligned}
$$

Proof of Proposition 6.6. The conditional prior odds are given by

$$
\begin{aligned}
CPONZ &= \frac{f(m_1 \mid \boldsymbol{\beta}_j = \mathbf{0}, \sigma^2)}{f(m_0 \mid \boldsymbol{\beta}_j = \mathbf{0}, \sigma^2)} \\
&= \frac{f(\boldsymbol{\beta}_j = \mathbf{0}|\sigma^2, m_1)f(\sigma^2|m_1)f(m_1)}{f(\boldsymbol{\beta}_j = \mathbf{0}|\sigma^2, m_0)f(\sigma^2|m_0)f(m_0)} \\
&= f(\boldsymbol{\beta}_j = \mathbf{0}|\sigma^2, m_1)\frac{f(\sigma^2|m_1)f(m_1)}{f(\sigma^2|m_0)f(m_0)} \\
&= f(\boldsymbol{\beta}_j = \mathbf{0}|\sigma^2, \gamma_j = 1)\frac{f(\sigma^2|m_1)f(m_1)}{f(\sigma^2|m_0)f(m_0)}. \triangleleft
\end{aligned}
$$

Proof of Proposition 6.7. The conditional prior odds are given by

$$
CPONZ = \frac{f(m_1 \mid \boldsymbol{\beta}_{(m_0)} = \mathbf{0}, \boldsymbol{\beta}_{(m_1)} = \mathbf{0}, \sigma^2)}{f(m_1 \mid \boldsymbol{\beta}_{(m_0)} = \mathbf{0}, \boldsymbol{\beta}_{(m_1)} = \mathbf{0}, \sigma^2)}
$$

[1]Since $f(\boldsymbol{\beta}_j^T\boldsymbol{\beta}_j < \xi^2|m_0) = f(\boldsymbol{\beta}_j = \mathbf{0}|m_0) = 1$

$$= \frac{f(\boldsymbol{\beta}_{(m_0)} = \mathbf{0}|\sigma^2, m_1) f(\boldsymbol{\beta}_{(m_1)} = \mathbf{0}|\sigma^2, m_1) f(\sigma^2|m_1) f(m_1)}{f(\boldsymbol{\beta}_{(m_0)} = \mathbf{0}|\sigma^2, m_0) f(\boldsymbol{\beta}_{(m_0)} = \mathbf{0}|\sigma^2, m_1) f(\sigma^2|m_0) f(m_0)}$$
$$= \frac{f(\boldsymbol{\beta}_{(m_1)} = \mathbf{0}|\sigma^2, m_1) f(\sigma^2|m_1) f(m_1)}{f(\boldsymbol{\beta}_{(m_0)} = \mathbf{0}|\sigma^2, m_0) f(\sigma^2|m_0) f(m_0)}. \triangleleft$$

Proof of Proposition 6.8. The proof is immediate if we substitute

$$\frac{f(m_1)}{f(m_0)} = \left(\frac{c^2}{\kappa}\right)^{[d(m_1)-d(m_0)]/2} \left(\frac{\left|\boldsymbol{V}_{(m_1)}\right| \left|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}^{-1}_{(m_1)}\right|}{\left|\boldsymbol{V}_{(m_0)}\right| \left|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}^{-1}_{(m_0)}\right|} \right)^{1/2}$$

in the penalty function (6.9) of proposition 6.1. $\triangleleft$

Proof of Corollary 6.8.1. The proof is immediate from the proposition 6.2 if we use the equation

$$2log\left(\frac{f(m_1)}{f(m_0)}\right) = \{d(m_1) - d(m_0)\} log(c^2+1) - \{d(m_1) - d(m_0)\} log(\kappa)$$

in the penalty function (6.11) of proposition 6.2. $\triangleleft$

Proof of Corollary 6.8.2. The prior model probability is given by

$$\begin{aligned}
2log\left(\frac{f(m_1)}{f(m_0)}\right) &= 2log\left(\frac{f(\boldsymbol{\gamma}_{m_1})}{f(\boldsymbol{\gamma}_{m_0})}\right) \\
&= 2\sum_{j\in\mathcal{V}} (\gamma_{j,m_1} - \gamma_{j,m_0}) log(\pi_j) - 2\sum_{j\in\mathcal{V}} (\gamma_{j,m_1} - \gamma_{j,m_0}) log(1-\pi_j) \\
&= 2\sum_{j\in\mathcal{V}} (\gamma_{j,m_1} - \gamma_{j,m_0}) log\left(\frac{\pi_j}{1-\pi_j}\right) \\
&= 2\sum_{j\in\mathcal{V}} (\gamma_{j,m_1} - \gamma_{j,m_0}) log\left(\frac{c^2+1}{\kappa}\right)^{d_j/2} \\
&= \sum_{j\in\mathcal{V}} (\gamma_{j,m_1} - \gamma_{j,m_0}) d_j log\left(\frac{c^2+1}{\kappa}\right) \\
&= \left(\sum_{j\in\mathcal{V}} \gamma_{j,m_1} d_j - \sum_{j\in\mathcal{V}} \gamma_{j,m_0} d_j\right) log\left(\frac{c^2+1}{\kappa}\right) \\
&= \{d(m_1) - d(m_0)\} log\left(\frac{c^2+1}{\kappa}\right).
\end{aligned}$$

If we use the above equality in the penalty function (6.11) we obtain the penalty function of corollary 6.8.2. $\triangleleft$

Proof of Corollary 6.8.3. The proof is immediate if we substitute

$$\boldsymbol{V}_{(m)} = D^{-1}_{(m)}(\boldsymbol{X}^T_j \boldsymbol{X}_j)$$

in the result of proposition 6.8. $\triangleleft$

Proof of Proposition 6.9. For the prior setup of proposition 6.9 the conditional prior odds at zero are given by

$$
\begin{aligned}
CPOZ &= \frac{f(\boldsymbol{\beta}_{(m_1)}|\sigma^2, m_1)f(\sigma^2|m_1)f(m_1)}{f(\boldsymbol{\beta}_{(m_0)}|\sigma^2, m_0)f(\sigma^2|m_0)f(m_0)}\\
&= (2\pi c^2\sigma^2)^{-[d(m_1)-d(m_0)]/2}\frac{|\boldsymbol{V}_{(m_1)}|^{-1/2}}{|\boldsymbol{V}_{(m_0)}|^{-1/2}}\\
&\times exp\left(\frac{1}{2c^2\sigma^2}\left(\boldsymbol{\mu}^T_{\beta_{(m_0)}}\boldsymbol{V}^{-1}_{(m_1)}\boldsymbol{\mu}_{\beta_{(m_1)}} - \boldsymbol{\mu}^T_{\beta_{(m_1)}}\boldsymbol{V}^{-1}_{(m_0)}\boldsymbol{\mu}_{\beta_{(m_0)}}\right)\right)\\
&\times \left(\frac{c^2}{\kappa}\right)^{[d(m_1)-d(m_0)]/2}\left(\frac{\left|\boldsymbol{V}_{(m_1)}\right|\left|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}^{-1}_{(m_1)}\right|}{\left|\boldsymbol{V}_{(m_0)}\right|\left|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}^{-1}_{(m_0)}\right|}\right)^{1/2}\\
&= (2\pi\kappa\sigma^2)^{-[d(m_1)-d(m_0)]/2}\left(\frac{\left|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}^{-1}_{(m_1)}\right|}{\left|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}^{-1}_{(m_0)}\right|}\right)^{1/2}\\
&\times exp\left(\frac{1}{2c^2\sigma^2}\left(\boldsymbol{\mu}^T_{\beta_{(m_0)}}\boldsymbol{V}^{-1}_{(m_1)}\boldsymbol{\mu}_{\beta_{(m_1)}} - \boldsymbol{\mu}^T_{\beta_{(m_1)}}\boldsymbol{V}^{-1}_{(m_0)}\boldsymbol{\mu}_{\beta_{(m_0)}}\right)\right). \triangleleft
\end{aligned}
$$

Proof of Corollary 6.9.1. For the prior setup of corollary 6.9.1 the conditional prior odds at zero are given by

$$
\begin{aligned}
CPOZ &= \frac{f(\boldsymbol{\beta}_{(m_1)}|\sigma^2, m_1)f(\sigma^2|m_1)f(m_1)}{f(\boldsymbol{\beta}_{(m_0)}|\sigma^2, m_0)f(\sigma^2|m_0)f(m_0)}\\
&= (2\pi c^2\sigma^2)^{-[d(m_1)-d(m_0)]/2}\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)}|^{1/2}}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}|^{1/2}}\left(\frac{c^2+1}{\kappa}\right)^{[d(m_1)-d(m_0)]/2}\\
&= (2\pi\sigma^2\kappa)^{-[d(m_1)-d(m_0)]/2}\left(\frac{c^2}{c^2+1}\right)^{-[d(m_1)-d(m_0)]/2}\left(\frac{|\boldsymbol{X}^T_{(m_1)}\boldsymbol{X}_{(m_1)}|}{|\boldsymbol{X}^T_{(m_0)}\boldsymbol{X}_{(m_0)}|}\right)^{1/2}. \triangleleft
\end{aligned}
$$

Proof of Corollary 6.9.2. The proof is immediate from proposition 6.9 substituting $\boldsymbol{V}_{(m)}$ by $D^{-1}_{(m)}(\boldsymbol{X}^T_j\boldsymbol{X}_j)$. $\triangleleft$

Proof of Proposition 6.10 . If we use the normal prior distribution $f(\mu|\sigma^2) \sim N(\mu_0, c^2\sigma^2)$ then the posterior odds are equal to

$$PO_{01} = \frac{f(m_0)}{1-f(m_0)}[\sqrt{nc^2+1}]exp\left(-\frac{n}{2\sigma^2}\frac{nc^2}{nc^2+1}(\bar{y}-\mu_0)^2\right).$$

The samples at the limit of significance satisfy the equality $\bar{y} = \mu_0 \pm z_{q/2}\sigma/\sqrt{n}$ which results in (6.31) if it is substituted in the above posterior odds. $\triangleleft$

Proof of Proposition 6.11. We substitute $f(m_0) = 1/(1+c)$ in the posterior odds at the limit of significance of lemma 6.10 and we obtain

$$POLS_{01}^{q} = c^{-1}\sqrt{nc^2+1}\; exp\left(-\frac{1}{2}\frac{nc^2}{nc^2+1}z_{q/2}^2\right)$$
$$= \sqrt{\frac{nc^2+1}{c^2}}\; exp\left(-\frac{1}{2}\frac{nc^2}{nc^2+1}z_{q/2}^2\right).$$

For large c^2 we have that

$$POLS_{01}^{q} \approx \sqrt{n}\; exp\left(-\frac{1}{2}z_{q/2}^2\right). \triangleleft$$

Proof of Proposition 6.12. The proof is immediate if we substitute

$$\frac{f(m_1)}{f(m_0)} = \left(\frac{c^2}{\kappa}\right)^{[d(m_1)-d(m_0)]/2}\left(\frac{\left|\boldsymbol{X}_{(m_1)}^T\boldsymbol{X}_{(m_1)} + c^{-2}D_{(m_1)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right|\prod_{j=1}^{p}|\boldsymbol{X}_j^T\boldsymbol{X}_j|^{-\gamma_{j,m_1}}}{\left|\boldsymbol{X}_{(m_0)}^T\boldsymbol{X}_{(m_0)} + c^{-2}D_{(m_0)}(\boldsymbol{X}_j^T\boldsymbol{X}_j)\right|\prod_{j=1}^{p}|\boldsymbol{X}_j^T\boldsymbol{X}_j|^{-\gamma_{j,m_0}}}\right)^{1/2}$$

in equation (6.32). $\triangleleft$

Proof of Proposition 6.13. We substitute (6.34) in the posterior odds at the limit of significance of proposition 6.10 and we obtain

$$POLS_{01}^{q} = \frac{\sqrt{nc^2+1}}{\sqrt{n}c}\frac{q}{1-q}\; exp\left(-\frac{1}{2}\frac{nc^2}{nc^2+1}z_{q/2}^2 + \frac{1}{2}z_{q/2}^2\right)$$
$$= \sqrt{\frac{nc^2+1}{nc^2}}\frac{q}{1-q}\; exp\left(\frac{1}{2}\frac{1}{nc^2+1}z_{q/2}^2\right).$$

For large nc^2 we have that

$$POLS_{01}^{q} \approx \frac{q}{1-q}. \triangleleft$$

Proof of Proposition 6.14. The proof is immediate if we substitute (6.35) in equation (6.32). $\triangleleft$

Proof of Proposition 6.15. Under the prior

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, m) \propto f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}_{(m)}^*, m)$$

the marginal likelihood is given by

$$f(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{X}_{(m)}^*, m) = \frac{\int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m)f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}^*, \boldsymbol{X}_{(m)}^*, m)d\boldsymbol{\beta}_{(m)}}{\int f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}^*, \boldsymbol{X}_{(m)}^*, m)d\boldsymbol{\beta}_{(m)}}$$
$$= \frac{\int f(\boldsymbol{y}, \boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, m)}{\int f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}^*, \boldsymbol{X}_{(m)}^*, m)d\boldsymbol{\beta}_{(m)}}.$$

Applying the Laplace approximation in both the numerator and the denominator of the above quantity we have the result of proposition 6.15. $\triangleleft$

Proof of Proposition 6.16. Under the prior

$$f(\boldsymbol{\beta}_{(m)}|\boldsymbol{y}^*, m) \propto [f(\boldsymbol{y}^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}^*_{(m)}, m)]^{1/c_0^2}$$

the marginal likelihood is given by

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{y}^*, \boldsymbol{X}^*_{(m)}, m) &= \frac{\int f(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m)[f(\boldsymbol{y}^*|\boldsymbol{\beta}^*_{(m)}, \boldsymbol{X}^*_{(m)}, m)]^{1/c_0^2} d\boldsymbol{\beta}_{(m)}}{\int [f(\boldsymbol{y}^*|\boldsymbol{\beta}^*_{(m)}, \boldsymbol{X}^*_{(m)}, m)]^{1/c_0^2} d\boldsymbol{\beta}_{(m)}} \\ &= \frac{\int \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}_{(m)}, m) \prod_{i=1}^{n_0} [f(y_i^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}_{(m)}, m)]^{1/c_0^2} d\boldsymbol{\beta}_{(m)}}{\int \prod_{i=1}^{n_0} [f(y_i^*|\boldsymbol{\beta}_{(m)}, \boldsymbol{X}_{(m)}, m)]^{1/c_0^2} d\boldsymbol{\beta}_{(m)}} \\ &= \frac{\int l(\boldsymbol{y}, \boldsymbol{y}^*, 1, c_0^{-2}|m) d\boldsymbol{\beta}_{(m)}}{\int l(\boldsymbol{y}, \boldsymbol{y}^*, 0, c_0^{-2}|m) d\boldsymbol{\beta}_{(m)}}. \end{aligned}$$

Applying the Laplace approximation in both the numerator and the denominator of the above quantity we have the result of proposition 6.16. $\triangleleft$

Proof of Proposition 6.17. In generalised linear models we have that

$$\left[\frac{\partial^2 log\left[f\left(\boldsymbol{y}|\boldsymbol{\beta}_{(m)}, m\right)\right]}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\hat{\boldsymbol{\beta}}_{(m)}} = \boldsymbol{X}^T_{(m)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m)}$$

where $\boldsymbol{H}_{(m)} = Diag(h_i)$, $h_i = \{g'(E[Y_i])^2 a_i(\phi) v(E[Y_i])\}^{-1}$ (see McCullagh and Nelder, 1983, for details). Using normal prior distribution $f\left(\boldsymbol{\beta}_{(m)}|m\right) \sim N(0, c^2\boldsymbol{V}_{(m)})$ for model parameters then, for large c^2 we have

$$|\mathcal{I}_{\breve{\boldsymbol{\beta}}_{(m)}}| = -\left[\frac{\partial^2 f_{m,y}(\boldsymbol{\beta}_{(m)})}{\partial\beta_{i,(m)}\partial\beta_{j,(m)}}\right]^{-1}_{\boldsymbol{\beta}_{(m)}=\breve{\boldsymbol{\beta}}_{(m)}} \approx \left(\boldsymbol{X}^T_{(m)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{V}^{-1}_{(m)}\right)^{-1}.$$

The normalising constants are given by

$$C[f(\boldsymbol{\beta}_{(m)}|m)] = (2\pi c^2)^{d(m)/2}|\boldsymbol{V}_{(m)}|^{1/2}.$$

Substituting the above two equalities in the penalty of equation (6.39) we have the penalty of proposition 6.17. $\triangleleft$

Proof of Corollary 6.17.1. The proof is immediate from proposition 6.17 if we substitute $\boldsymbol{V}_{(m)}$ by $\left(\boldsymbol{X}^T_{(m)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m)}\right)^{-1}$. $\triangleleft$

Proof of Proposition 6.18. The proof is immediate if we substitute

$$\frac{f(m_1)}{f(m_0)} = \left(\frac{c^2}{\kappa}\right)^{[d(m_1)-d(m_0)]/2} \left(\frac{|\boldsymbol{V}_{(m_1)}||\boldsymbol{X}^T_{(m_1)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m_1)} + c^{-2}\boldsymbol{V}_{(m_1)}|}{|\boldsymbol{V}_{(m_0)}||\boldsymbol{X}^T_{(m_0)}\boldsymbol{H}_{(m)}\boldsymbol{X}_{(m_0)} + c^{-2}\boldsymbol{V}_{(m_0)}|}\right)^{1/2}$$

in the penalty function of proposition 6.17. $\triangleleft$

Proof of Corollary 6.18.1. The proof is immediate if we substitute

$$\frac{f(m_1)}{f(m_0)} = \left(\frac{c^2+1}{\kappa}\right)^{[d(m_1)-d(m_0)]/2}$$

in the penalty function of corollary 6.17.1. $\triangleleft$

Proof of Proposition 6.19. Without loss of generality, for every model m we can write

$$\boldsymbol{X} = [\boldsymbol{X}_{(m)}\boldsymbol{X}_{(\backslash m)}], \quad \boldsymbol{\beta}^T = [\boldsymbol{\beta}^T_{(m)}\boldsymbol{\beta}^T_{(\backslash m)}], \quad \boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_{(m)} & \boldsymbol{R}_{(m,\backslash m)} \\ \boldsymbol{R}^T_{(m,\backslash m)} & \boldsymbol{R}_{(\backslash m)} \end{bmatrix}$$

where $\boldsymbol{X}_{(\backslash m)}$ and $\boldsymbol{\beta}_{(\backslash m)}$ refer to the components of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ excluded from model m. The matrix $\boldsymbol{R}$ is partitioned to matrices: $\boldsymbol{R}_{(m)}$ that corresponds to covariances between terms included in model m; $\boldsymbol{R}_{(\backslash m)}$ that corresponds to covariances between terms excluded from model m; and $\boldsymbol{R}_{(m,\backslash m)}$ that corresponds to covariances between each term included in model m and each term excluded from model m. In such case the prior matrix is given by

$$\boldsymbol{V}^{SSVS}_{(m)} = \begin{bmatrix} \boldsymbol{R}_{(m)} & k^{-1}\boldsymbol{R}_{(m,\backslash m)} \\ k^{-1}\boldsymbol{R}^T_{(m,\backslash m)} & k^{-2}\boldsymbol{R}_{(\backslash m)} \end{bmatrix}$$

$$[\boldsymbol{V}^{SSVS}_{(m)}]^{-1} = \begin{bmatrix} \boldsymbol{R}^-_{(m)} & -k\boldsymbol{R}^-_{(m,\backslash m)} \\ -k\boldsymbol{R}^-_{(\backslash m,m)} & k^2\boldsymbol{R}^-_{(\backslash m)} \end{bmatrix}$$

$$\begin{aligned}
\boldsymbol{R}^-_{(m)} &= [\boldsymbol{R}_{(m)} - \boldsymbol{R}_{(m,\backslash m)}\boldsymbol{R}^{-1}_{(\backslash m)}\boldsymbol{R}^T_{(m,\backslash m)}]^{-1}, \\
\boldsymbol{R}^-_{(\backslash m)} &= \boldsymbol{R}^{-1}_{(\backslash m)} + \boldsymbol{R}^{-1}_{(\backslash m)}\boldsymbol{R}^T_{(m,\backslash m)}\boldsymbol{R}^-_{(m)}\boldsymbol{R}_{(m,\backslash m)}\boldsymbol{R}^{-1}_{(\backslash m)} \\
&= [\boldsymbol{R}_{(\backslash m)} - \boldsymbol{R}_{(\backslash m,m)}\boldsymbol{R}^{-1}_{(m)}\boldsymbol{R}^T_{(\backslash m,m)}]^{-1}, \\
\boldsymbol{R}^-_{(m,\backslash m)} &= \boldsymbol{R}^-_{(m)}\boldsymbol{R}_{(m,\backslash m)}\boldsymbol{R}^{-1}_{(\backslash m)} \\
&= \boldsymbol{R}^{-1}_{(m)}\boldsymbol{R}_{(m,\backslash m)}[\boldsymbol{R}_{(\backslash m)} - \boldsymbol{R}_{(\backslash m,m)}\boldsymbol{R}^{-1}_{(m)}\boldsymbol{R}^T_{(\backslash m,m)}]^{-1}, \\
\boldsymbol{R}^-_{(\backslash m,m)} &= \boldsymbol{R}^{-1}_{(\backslash m)}\boldsymbol{R}^T_{(m,\backslash m)}\boldsymbol{R}^-_{(m)} \\
&= \boldsymbol{R}^{-1}_{(\backslash m)}\boldsymbol{R}_{(\backslash m,m)}[\boldsymbol{R}_{(m)} - \boldsymbol{R}_{(m,\backslash m)}\boldsymbol{R}^{-1}_{(\backslash m)}\boldsymbol{R}^T_{(m,\backslash m)}]^{-1}.
\end{aligned}$$

Furthermore, the determinant is given by

$$|\boldsymbol{V}^{SSVS}_{(m)}|^{-1} = [k^2]^{d-d(m)}|\boldsymbol{R}^-_{(\backslash m)}||\boldsymbol{R}^-_{(m)} - \boldsymbol{R}^-_{(m,\backslash m)}[\boldsymbol{R}^-_{(\backslash m)}]^{-1}\boldsymbol{R}^-_{(\backslash m,m)}|.$$

The posterior odds using the SSVS approach are given by

$$PO_{01} = \left(\frac{|\boldsymbol{V}^{SSVS}_{(m_0)}|}{|\boldsymbol{V}^{SSVS}_{(m_1)}|} \right)^{-1/2} \left(\frac{|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}^{SSVS}_{(m_0)}]^{-1}|}{|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}^{SSVS}_{(m_1)}]^{-1}|} \right)^{1/2} \left(\frac{SS^{SSVS}_{m_0}}{SS^{SSVS}_{m_1}} \right)^{-n/2} \frac{f(m_0)}{f(m_1)}$$

with

$$SS^{SSVS}_m = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X} \left(\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}^{SSVS}_{(m)}]^{-1} \right)^{-1} \boldsymbol{X}^T\boldsymbol{y}$$

The $\left(\boldsymbol{X}^T\boldsymbol{X} + c^{-2}\boldsymbol{V}^{SSVS}_{(m)} \right)^{-1}$ is given by

$$\begin{aligned}
\left(\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}^{SSVS}_{(m)}]^{-1} \right)^{-1} &= \begin{bmatrix} \boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}^-_{(m)} & \boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k\boldsymbol{R}^-_{(m,\backslash m)} \\ \boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(m)} + c^{-2}k\boldsymbol{R}^-_{(\backslash m,m)} & \boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}^-_{(m)} & \tilde{\boldsymbol{\Sigma}}^-_{(m,\backslash m)} \\ \tilde{\boldsymbol{\Sigma}}^-_{(\backslash m,m)} & \tilde{\boldsymbol{\Sigma}}^-_{(\backslash m)} \end{bmatrix}
\end{aligned}$$

with

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}^-_{(m)} &= \Big[\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}^-_{(m)} - [\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k\boldsymbol{R}^-_{(m,\backslash m)}] \times \\
&\quad \times [\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)}]^{-1}[\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(m)} + c^{-2}k\boldsymbol{R}^-_{(\backslash m,m)}] \Big]^{-1}, \\
\tilde{\boldsymbol{\Sigma}}^-_{(\backslash m)} &= [\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)}]^{-1} + \\
&\quad + [\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)}]^{-1}[\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(m)} + c^{-2}k\boldsymbol{R}^-_{(\backslash m,m)}] \times \\
&\quad \times \tilde{\boldsymbol{\Sigma}}^-_{(m)}[\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k\boldsymbol{R}^-_{(m,\backslash m)}][\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)}]^{-1}, \\
\tilde{\boldsymbol{\Sigma}}^-_{(m,\backslash m)} &= -\tilde{\boldsymbol{\Sigma}}^-_{(m)} \left(\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k\boldsymbol{R}^-_{(m,\backslash m)} \right) \left(\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)} \right)^{-1}, \\
\tilde{\boldsymbol{\Sigma}}^-_{(\backslash m,m)} &= -\left(\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}^-_{(\backslash m)} \right)^{-1} \left(\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(m)} + c^{-2}k\boldsymbol{R}^-_{(\backslash m,m)} \right) \tilde{\boldsymbol{\Sigma}}^-_{(m)}.
\end{aligned}$$

$$\begin{aligned}
\tilde{\boldsymbol{\Sigma}}^-_{(m)} &= \Big[\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}^-_{(m)} - [k^{-1}\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}^-_{(m,\backslash m)}] \times \\
&\quad \times [k^{-2}\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}^-_{(\backslash m)}]^{-1}[k^{-1}\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}^-_{(\backslash m,m)}] \Big]^{-1} \\
\lim_{l\to\infty} \tilde{\boldsymbol{\Sigma}}^-_{(m)} &= \Big[\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}^-_{(m)} - c^{-2}\boldsymbol{R}^-_{(m,\backslash m)}[\boldsymbol{R}^-_{(\backslash m)}]^{-1}\boldsymbol{R}^-_{(\backslash m,m)} \Big]^{-1}, \\
\tilde{\boldsymbol{\Sigma}}^-_{(\backslash m)} &= k^{-2}[k^{-2}\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}^-_{(\backslash m)}]^{-1} + k^{-2}[k^{-2}\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}^-_{(\backslash m)}]^{-1} \times \\
&\quad \times [k^{-1}\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}^-_{(\backslash m,m)}]\tilde{\boldsymbol{\Sigma}}^-_{(m)}[k^{-1}\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}^-_{(m,\backslash m)}] \times \\
&\quad \times [k^{-2}\boldsymbol{X}^T_{(\backslash m)}\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}^-_{(\backslash m)}]^{-1}, \\
\lim_{k\to\infty} \tilde{\boldsymbol{\Sigma}}^-_{(\backslash m)} &= \boldsymbol{0}, \\
\lim_{k\to\infty} \tilde{\boldsymbol{\Sigma}}^-_{(m,\backslash m)} &= \boldsymbol{0}, \\
\lim_{k\to\infty} \tilde{\boldsymbol{\Sigma}}^-_{(\backslash m,m)} &= \boldsymbol{0}.
\end{aligned}$$

From the above we have

$$\begin{aligned}\lim_{k\to\infty} SS_m^{SSVS} &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}_{(m)}\left[\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)} + c^{-2}[\boldsymbol{R}_{(m)}^- - \boldsymbol{R}_{(m,\backslash m)}^-[\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-]\right]^{-1}\boldsymbol{X}^T\boldsymbol{y}\\ &= SS_m.\end{aligned}$$

The determinant is given by

$$\begin{aligned}|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}\boldsymbol{V}_{(m)}| &= |\tilde{\boldsymbol{\Sigma}}_{(m)}^-||\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^- - \tilde{\boldsymbol{\Sigma}}_{(\backslash m,m)}^-[\tilde{\boldsymbol{\Sigma}}_{(m)}^-]^{-1}\tilde{\boldsymbol{\Sigma}}_{(m,\backslash m)}^-|\\ &= |\tilde{\boldsymbol{\Sigma}}_{(m)}^-||\boldsymbol{I} - \tilde{\boldsymbol{\Sigma}}_{(\backslash m,m)}^-[\tilde{\boldsymbol{\Sigma}}_{(m)}^-]^{-1}\tilde{\boldsymbol{\Sigma}}_{(m,\backslash m)}^-[\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-]^{-1}||\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-|\end{aligned}$$

$$\begin{aligned}|\boldsymbol{V}_{(m)}^{SSVS}|^{-1}|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}\boldsymbol{V}_{(m)}| &= |\boldsymbol{R}_{(\backslash m)}^-||\boldsymbol{R}_{(m)}^- - \boldsymbol{R}_{(m,\backslash m)}^-[\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-|\\ &\quad\times|\tilde{\boldsymbol{\Sigma}}_{(m)}^-||\boldsymbol{I} - \tilde{\boldsymbol{\Sigma}}_{(\backslash m,m)}^-[\tilde{\boldsymbol{\Sigma}}_{(m)}^-]^{-1}\tilde{\boldsymbol{\Sigma}}_{(m,\backslash m)}^-[\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-]^{-1}||k^2\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-|\end{aligned}$$

$$\begin{aligned}\lim_{k\to\infty}\left(k^2\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-\right) &= \lim_{k\to\infty}\Big([k^{-2}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(\backslash m)}^-]^{-1} + [k^{-2}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(\backslash m)}^-]^{-1}\times\\ &\quad\times[k^{-1}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}_{(\backslash m,m)}^-]\tilde{\boldsymbol{\Sigma}}_{(m)}^-[k^{-1}\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(m,\backslash m)}^-]\times\\ &\quad\times[k^{-2}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(\backslash m)}^-]^{-1}\Big)\\ &= c^2[\boldsymbol{R}_{(\backslash m)}^-]^{-1} + [\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-\\ &\quad\times\left[\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}_{(m)}^- - c^{-2}\boldsymbol{R}_{(m,\backslash m)}^-[\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-\right]^{-1}\boldsymbol{R}_{(m,\backslash m)}^-[\boldsymbol{R}_{(\backslash m)}^-]^{-1}\Big)\\ &= c^2[\boldsymbol{R}_{(\backslash m)}^-]^{-1} + \boldsymbol{R}^*,\end{aligned}$$

where $\boldsymbol{R}^*$ is given by

$$\boldsymbol{R}^* = [\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-\left[\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}_{(m)}^- - c^{-2}\boldsymbol{R}_{(m,\backslash m)}^-[\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-\right]^{-1}\boldsymbol{R}_{(m,\backslash m)}^-[\boldsymbol{R}_{(\backslash m)}^-]^{-1}.$$

Consider now the limit

$$\lim_{k\to\infty}\left(\tilde{\boldsymbol{\Sigma}}_{(\backslash m,m)}^-[\tilde{\boldsymbol{\Sigma}}_{(m)}^-]^{-1}\tilde{\boldsymbol{\Sigma}}_{(m,\backslash m)}^-[\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-]^{-1}\right) =$$

$$\begin{aligned}&= \lim_{k\to\infty}\left(\left(\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}_{(\backslash m)}^-\right)^{-1}\left(\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(m)} + c^{-2}k\boldsymbol{R}_{(\backslash m,m)}^-\right)\tilde{\boldsymbol{\Sigma}}_{(m)}^-\right.\\ &\quad\left.\times\left(\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}k\boldsymbol{R}_{(m,\backslash m)}^-\right)\left(\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}k^2\boldsymbol{R}_{(\backslash m)}^-\right)^{-1}[\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-]^{-1}\right)\\ &= \lim_{k\to\infty}\left(\left(k^{-2}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(\backslash m)}^-\right)^{-1}\left(k^{-1}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(m)} + c^{-2}\boldsymbol{R}_{(\backslash m,m)}^-\right)\tilde{\boldsymbol{\Sigma}}_{(m)}^-\right.\\ &\quad\left.\times\left(k^{-1}\boldsymbol{X}_{(m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(m,\backslash m)}^-\right)\left(k^{-2}\boldsymbol{X}_{(\backslash m)}^T\boldsymbol{X}_{(\backslash m)} + c^{-2}\boldsymbol{R}_{(\backslash m)}^-\right)^{-1}[k^2\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-]^{-1}\right)\\ &= [\boldsymbol{R}_{(\backslash m)}^-]^{-1}\boldsymbol{R}_{(\backslash m,m)}^-\lim_{k\to\infty}\left(\tilde{\boldsymbol{\Sigma}}_{(m)}^-\right)\boldsymbol{R}_{(m,\backslash m)}^-\boldsymbol{R}_{(\backslash m)}^-\lim_{k\to\infty}\left[k^2\tilde{\boldsymbol{\Sigma}}_{(\backslash m)}^-\right]^{-1}\\ &= \boldsymbol{R}^*\left(c^2[\boldsymbol{R}_{(\backslash m)}^-]^{-1} + \boldsymbol{R}^*\right)^{-1}.\end{aligned}$$

Therefore, from the above have

$$\lim_{k\to\infty} |\boldsymbol{V}_{(m)}^{SSVS}|^{-1}|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}\boldsymbol{V}_{(m)}^{SSVS}| \;\; =$$

$$\begin{aligned}
&= |\boldsymbol{R}^{-}_{(\backslash m)}||\boldsymbol{R}^{-}_{(m)} - \boldsymbol{R}^{-}_{(m,\backslash m)}[\boldsymbol{R}^{-}_{(\backslash m)}]^{-1}\boldsymbol{R}^{-}_{(\backslash m,m)}| \\
&\quad \times \left|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} + c^{-2}[\boldsymbol{R}^{-}_{(m)} - \boldsymbol{R}^{-}_{(m,\backslash m)}[\boldsymbol{R}^{-}_{(\backslash m)}]^{-1}\boldsymbol{R}^{-}_{(\backslash m,m)}]\right|^{-1} |\tilde{\boldsymbol{\Sigma}}^{-}_{(m)}| \\
&= [c^2]^{d-d(m)}|\boldsymbol{R}^{-}_{(\backslash m)}||\boldsymbol{R}^{-}_{(m)} - \boldsymbol{R}^{-}_{(m,\backslash m)}[\boldsymbol{R}^{-}_{(\backslash m)}]^{-1}\boldsymbol{R}^{-}_{(\backslash m,m)}||\tilde{\boldsymbol{\Sigma}}^{-}_{(m)}||\boldsymbol{R}^{-}_{(\backslash m)}|^{-1} \\
&= [c^2]^{d-d(m)}|\boldsymbol{V}_{(m)}|^{-1}|\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)} - c^{-2}\boldsymbol{V}^{-1}_{(m)}|
\end{aligned}$$

with

$$\boldsymbol{V}^{-1}_{(m)} = \boldsymbol{R}^{-}_{(m)} - \boldsymbol{R}^{-}_{(m,\backslash m)}[\boldsymbol{R}^{-}_{(\backslash m)}]^{-1}\boldsymbol{R}^{-}_{(\backslash m,m)} = \boldsymbol{R}^{-1}_{(m)}.$$

From the above result and the fact that $SS_m^{SSVS} \to SS_m$, for large k^2, we obtain the statement of proposition 6.19. $\triangleleft$

Proof of Proposition 6.20. In linear regression the Fisher information matrix, $\mathcal{I}^{-1}_{\boldsymbol{\beta}_{(m)}}$ of model m is given by $\boldsymbol{X}^T_{(m)}\boldsymbol{X}_{(m)}\sigma^{-2}$. Standard asymptotic theory requires a regularity condition in which $\frac{1}{n}\mathcal{I}^{-1}_{\boldsymbol{\beta}_{(m)}}$ has a positive definite limit. Therefore, under mild regularity conditions

$$\frac{|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}^{SSVS}_{(m_1)}]^{-1}|}{|\boldsymbol{X}^T\boldsymbol{X} + c^{-2}[\boldsymbol{V}^{SSVS}_{(m_0)}]^{-1}|} = \frac{|n^{-1}\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + c^{-2}n^{-1}\sigma^{-2}[\boldsymbol{V}^{SSVS}_{(m_1)}]^{-1}|}{|n^{-1}\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + c^{-2}n^{-1}\sigma^{-2}[\boldsymbol{V}^{SSVS}_{(m_0)}]^{-1}|} \to 1.$$

From the above it is also direct that for $n \to \infty$ and fixed prior variances $\boldsymbol{V}^{SSVS}_{(m)}$ then

$$SS_m^{SSVS} = RSS_{full}$$

resulting to the proposition 6.20. $\triangleleft$

Proof of Proposition 6.21. The proof is direct if we consider that

$$\lim_{c^2\to\infty}\left(\boldsymbol{X}^T\boldsymbol{X} + c^{-2}\boldsymbol{V}^{SSVS}_{(m)}\right) = \boldsymbol{X}^T\boldsymbol{X}$$

and therefore

$$\lim_{c^2\to\infty}(SS_m^{SSVS}) = RSS_{full}$$

where RSS_{full} is the residual sum of squares of the full model. Substituting the above limits in the SSVS based posterior odds we have the result of proposition 6.21. $\triangleleft$

Chapter 7

Gibbs Variable Selection Using Bugs

The aim of this chapter is to clearly illustrate how we can utilize BUGS (Spiegelhalter *et al.*, 1996a) for the implementation of variable selection methods. We concentrate on Gibbs variable selection, proposed in Chapter 4, using independent prior distributions. Extension to SSVS and Kuo and Mallick samplers is straightforward. Note that this chapter is also given in a form of research paper; see Ntzoufras (1999b).

7.1 Definition of likelihood

The likelihood (3.2) used in Gibbs variable selection and Kuo and Mallick sampler can be easily incorporated in BUGS using the following code

```
for (i in 1:N) { for (j in 1:p) { z[i,j]<-x[i,j]*b[j]*g[j]}}
for (i in 1:N) { eta[i] <-sum(z[i,]) ;
                 y[i]~distribution [ parameter1, parameter2 ]  }
```

where

- `N` denotes the sample size,
- `p` the number of total variables under consideration,
- `x[i,j]` is the i, j component of the data or design matrix $\boldsymbol{X}$,
- `y[i]` is i element of the response vector $\boldsymbol{y}$,

- `b[j]` is the j element of the parameter vector $\boldsymbol{\beta}$,
- `g[j]` is the inclusion indicator for j element of $\boldsymbol{\gamma}$,
- `z[i,j]` is a matrix which is used to simplify calculations,
- `eta[i]` is the i element of linear predictor vector $\boldsymbol{\eta}$ and should be substituted by the corresponding link function, for example `logit(p[i])` in binomial logistic regression,
- `distribution` should be substituted by appropriate BUGS command for the distribution that the user wants to use (for example `dnorm` for normal distribution),
- `parameter1,parameter2` should be substituted according to distribution chosen, for example for the normal distribution with mean μ_i and variance τ^{-1} we may use `mu[i], tau`.

Alternatively, if we prefer to use SSVS as defined by George and McCulloch (1993) should change the first line of the above code to

```
for (i in 1:N) { for (j in 1:p) { z[i,j]<-x[i,j]*b[j]}}.
```

For the usual normal, binomial and Poisson models the model formulations are given by the following lines of BUGS code

Normal: `for (i in 1:N) { mu[i] <-sum(z[i,]) ; y[i]~dnorm (mu[i], tau)  }`

where `mu[i]` is the expected value for the ith observation and `tau` is the precision of the regression model.

Poisson:
```
for (i in 1:N) { log(lambda[i])   <- sum(z[i,]);
                 y[i] ~ dpois(lambda[i])}
```

where `lambda[i]` is the Poisson mean for the ith observation.

Binomial:
```
for (i in 1:N) { logit(p[i])   <- sum(z[i,]);
                 y[i] ~ dbin(p[i], n[i])}
```

where `p[i]` is the probability of success and `n[i]` is the total number of Bernoulli trials for the ith binomial experiment. Alternative link functions maybe used by substituting `logit(p[i])` by `probit(p[i])` or `cloglog(p[i])` for $\Phi^{-1}(p)$ and $log(-log(1-p))$; where Φ is the normal cumulative distribution function.

7.2 Definition of Prior Distribution of β

In situations where we use independent priors similar to (4.4) and each covariate parameter vector is univariate, the definition of the prior is straightforward. Our prior is a mixture of independent normal distributions

$$\beta_j \sim (1-\gamma_j)N(\bar{\mu}_j, S_j) + \gamma_j N(0, \Sigma_j), \quad j = 1, 2, \dots, p \tag{7.1}$$

where $\bar{\mu}_j$, S_j are the mean and variance respectively, used in the corresponding pseudoprior distributions and Σ_j is the prior variance, when the j term is included in the model. In order to use (7.1) in BUGS we write

- `b[j]` $\sim$ `dnorm( bpriorm[j], tprior[j])` denoting $\beta_j \sim N(m_j, \tau_j^{-1})$,
- `bpriorm[j]` $<-$ `(1-g[j])*mean[j]` denoting $m_j = (1-\gamma_j)\bar{\mu}_j$,
- `tprior[j]` $<-$ `g[j]*t[j]+(1-g[j])*pow(se[j],-2)` denoting $\tau_j = (1-\gamma_j)S_j^{-1} + \gamma_j \Sigma_j^{-1}$,

for $j = 1, 2, \dots, p$; where m_j and τ_j are the prior mean and precision for β_j depending on γ_j and `t[j]`, `se[j]`, `mean[j]`, `bpriorm[j]`, `tprior[j]` are the BUGS variables for Σ_j^{-1}, $\sqrt{S_j}$, $\bar{\mu}_j$, m_j and τ_j, respectively.

When we have multivariate $\boldsymbol{\beta}_j$ then the vector $\boldsymbol{\beta}$ has greater dimensionality than $\boldsymbol{\gamma}$. In these situations we denote by p and $d(>p)$ the dimensions of $\boldsymbol{\gamma}$ and the full parameter vector $\boldsymbol{\beta}$, respectively. Therefore, we need one variable to facilitate the association between these two vectors. This vector is denoted by the BUGS variable `pos`. The `pos` vector, which has dimension equal to the dimension of $\boldsymbol{\beta}$, takes values from $1, 2, ..., p$ and denotes that β_k is related to γ_{pos_k} for $k = 1, 2, ..., d$.

Here we illustrate the use of a mixture of normal prior distributions as in (4.4). This prior can be expressed as a multivariate normal distribution on the 'full' parameter vector $\boldsymbol{\beta}$. Therefore we write in BUGS

- `b[ ]` $\sim$ `dmnorm( bpriorm[ ], Tau[,])` denoting $\boldsymbol{\beta} \sim N_d(\boldsymbol{m}, \boldsymbol{T}^{-1})$,
- `bpriorm[k]` $<-$ `(1-g[pos[k]])*mean[k]` denoting $m_k = (1-\gamma_{pos_k})\bar{\mu}_k$,

- `Tau[k,l] <- (1-g[pos[k]]*g[pos[l]])*t[k,l]+`
  `+ g[pos[k]]*gamma[pos[l]]*equals(k,l)*pow(se[k],-2)` denoting that

$$T_{kl} = \begin{cases} [\boldsymbol{\Sigma}^{-1}]_{kl} & \text{when } \gamma_{pos_k} = \gamma_{pos_l} = 1 \\ se_k^{-2} & \text{when } k = l \ \&\ \gamma_{pos_k} = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{for } k, l = 1, 2, \ldots, d;$$

where N_d is the d-dimensional normal distribution; $\mathbf{m}^T = (m_1, m_2, \ldots, m_d)$ and $\mathbf{T}$ are the prior mean vector and precision matrix depending on $\boldsymbol{\gamma}$; $\bar{\mu}_k$ is the corresponding pilot run estimate for k element of model parameter vector $\boldsymbol{\beta}$; $\boldsymbol{\Sigma}$ is the constructed prior covariance matrix for the whole parameter vector $\boldsymbol{\beta}$ when we use for each $\boldsymbol{\beta}_j$ the multivariate extension of prior distribution 7.1; T_{kl} and $[\boldsymbol{\Sigma}^{-1}]_{kl}$ is the k row and l column elements of $\boldsymbol{T}$ and $\boldsymbol{\Sigma}^{-1}$ matrices respectively; and `Tau[,]`, `t[,]` are the BUGS matrices for $\boldsymbol{T}$ and $\boldsymbol{\Sigma}^{-1}$, respectively. For application of the above see example 1.

SSVS and Kuo Mallick sampler can be easily applied by slightly changing the above code. In SSVS the prior (7.1) is used with $\bar{\mu}_j = 0$ and $S_j = \Sigma_j / k_j^2$, where k_j^2 should be large enough in order that β_j will be close to zero when $\gamma_j = 0$. For selection of the prior parameters in SSVS see semiautomatic prior selection of George and McCulloch (1993, 1997). The above restriction can be easily applied in BUGS by

```
bpriorm[j] <- 0
tprior[j] <- t[j]*g[j]+(1-g[j])*t[j]*pow(k[j],2) .
```

Kuo and Mallick sampler uses prior on $\boldsymbol{\beta}$ that does not depend on model indicator $\boldsymbol{\gamma}$. Therefore prior specification is the same as in simple modelling with BUGS; for more details see Spiegelhalter *et al.* (1996a,b,c).

7.3 Definition of Prior Term Probabilities

In order to apply any variable selection method in BUGS we need to define the prior probabilities $f(\boldsymbol{\gamma})$. When we are vague about models we may set $f(\boldsymbol{\gamma}) = 1/|\mathcal{M}|$, where $|\mathcal{M}|$ is the number of all models under consideration. When the explanatory variables do not involve interactions (e.g. linear regression) then the number of models under consideration is 2^p. In these situations the latent variables γ_j can be treated as $a-priori$ independent and therefore set in BUGS

- `g[j]` $\sim$ `dbern(0.5)` denoting that $\gamma_j \sim Bernoulli(0.5)$.

for all $j = 1, 2, \ldots, p$. This prior results to $f(\boldsymbol{\gamma}) = 2^{-p} \; \forall \; \boldsymbol{\gamma} \in \{0,1\}^p$. When we are dealing with models using categorical explanatory variables with interaction terms, such as $ANOVA$ or log-linear models, we usually want to restrict attention to hierarchical models. The conditional distributions of $f(\gamma_j|\boldsymbol{\gamma}_{\setminus j})$ need to be specified in such way that $f(\boldsymbol{\gamma}) = |\mathcal{M}|^{-1}$ when $\boldsymbol{\gamma}$ is referring to hierarchical model and $f(\boldsymbol{\gamma}) = 0$ otherwise.

For example, in a two way $ANOVA$ we have three terms under consideration ($p = 3$). All possible models are eight, while the plausible models are only five ($constant, [A], [B], [A][B]$ and $[AB]$). Therefore, we need to have $f(\boldsymbol{\gamma}) = 0.20$ for the above five models and $f(\boldsymbol{\gamma}) = 0$ for the rest. This can be applied by setting in BUGS

- `g[3]` $\sim$ `dbern(0.2)` denoting that $\gamma_{AB} \sim Bernoulli(0.2)$.
- `pi` $<-$ `g[3]+0.5(1-g[3])` denoting that $\pi = \gamma_{AB} + 0.5(1 - \gamma_{AB})$,
- `for (i in 1:2) {` `g[j]` $\sim$ `dbern(pi) }` denoting that $\forall \; i \in \{A, B\}$, $\gamma_j|\gamma_{AB} \sim Bernoulli(\pi)$

From the above it is evident that

$$f([AB]) = f(\gamma_{AB} = 1)f(\gamma_A = 1|\gamma_{AB} = 1)f(\gamma_B = 1|\gamma_{AB} = 1) = 0.2 \times 1 \times 1 = 0.2$$

$$f([A][B]) = f(\gamma_{AB} = 0)f(\gamma_A = 1|\gamma_{AB} = 0)f(\gamma_B = 1|\gamma_{AB} = 0) = 0.8 \times 0.5 \times 0.5 = 0.2$$

Using similar calculations we find that $f(\boldsymbol{\gamma}) = 0.2$ for all five models under consideration. For further relevant discussion and application see Chipman (1996). For implementation in BUGS see examples 1 and 4.

7.4 Calculating Model Probabilities in Bugs

In order to directly calculate in BUGS the posterior model probabilities and avoid saving large output we can use matrix type variables with dimension equal to the number of models.

Using a simple coding such as $1+\sum_{j=1}^{p}\gamma_j 2^{j-1}$ we transform the vector $\boldsymbol{\gamma}$ in a unique, for each model index (noted by `mdl`) for which `pmdl[mdl]=1` and `pmdl[j]=0` for all `j` $\neq$ `mdl`. The above statements can be written in BUGS with the code

```
for (j in 1:p) { index[j] <- pow(2,j-1) }
mdl <- 1+inprod(g[ ], index[ ])
for (m in 1:mdl) { pmdl[m] <- equals(m,mdl) }
```

Then using the command `stats(pmdl)` in BUGS environment (or cmd file) we can monitor the posterior model probabilities. This is feasible only if the number of models is limited and therefore applicable only in some simple cases.

7.5 Examples

The implementation of four illustrated examples are briefly presented. The first example is a $3\times 2\times 4$ contingency table presented in Section 4.6.2.4 used to illustrate how to handle factors with more than two levels. Example 2 is a logistic regression example in which we use the orthogonal transformed space of Clyde *et al.* (1996). Example 3 provides model selection details in a regression type problem involving many different error distributions while example 4 is a simple logistic regression problem with random effects. In all examples posterior probabilities are presented while the associated BUGS codes are provided in the appendix. Additional details (for example, convergence plots) are omitted since the aim of this chapter is only to illustrate how to use BUGS for variable selection.

7.5.1 Example 1: $3\times 2\times 4$ Contingency Table

This example was presented in Section 4.6.2.4. The BUGS results presented in Table 7.1 can be compared with the results of FORTRAN 77 code presented in Section 4.6.2.4. The full model is given by

$$n_{ilk}\sim Poisson(\lambda_{ilk}),\quad log(\lambda_{ilk})=m+o_i+h_l+a_k+oh_{il}+oa_{ik}+ha_{lk}+oha_{ilk},$$

for $i=1,2,3,\ l=1,2,\ k=1,2,3,4$. The above model can be rewritten with likelihood given by (3.2) where $\boldsymbol{\beta}$ can be divided to $\boldsymbol{\beta}_j$ sub-vectors with $j\in\{\emptyset,O,H,OH,A,OA,HA,OHA\}$; where $\beta_\emptyset=m$, $\boldsymbol{\beta}_O^T=[o_2,o_3]$, $\beta_H=h_2$, $\beta_{OH}^T=[oh_{22},oh_{32}]$, $\boldsymbol{\beta}_A^T=[a_2,a_3,a_4]$, $\boldsymbol{\beta}_{OA}^T=$

Pseudopriors	k=10		Pilot Run	
Burn-in	1,000	10,000	1,000	10,000
Iterations	1,000	$10 \times 10,000$	1,000	$10 \times 10,000$
<u>Models</u>				
$[O][H][A]$	62.80	68.87	65.20	67.80
$[OH][A]$	36.90	30.53	34.40	31.63
$[O][HA]$	0.20	0.40	0.10	0.43
$[OH][HA]$	0.10	0.20	0.30	0.14
<u>Terms</u>				
$\gamma_{OH} = 1$	37.00	30.63	34.70	31.77
$\gamma_{HA} = 1$	0.30	0.20	0.40	0.57

Table 7.1: $3 \times 2 \times 4$ Contingency Table: Posterior Model Probabilities Using BUGS.

$[oa_{22}, oa_{23}, oa_{32}, oa_{33}]$, $\boldsymbol{\beta}_{HA}^T = [ha_{22}, ha_{23}]$ and $\boldsymbol{\beta}_{OHA}^T = [oha_{222}, oha_{223}, oha_{322}, oha_{323}]$. Each $\boldsymbol{\beta}_j$ is a multivariate vector and therefore each prior distribution involves mixture multivariate normal distributions. We use sum to zero constraints and prior variance Σ_j as in Dellaportas and Forster (1999). We restrict attention in hierarchical models including always the main effects since we are mainly interested for relationships between the categorical factors. Under these restrictions the models under consideration are nine and in order to forbid to move to non hierarchical models we use the following priors in BUGS

- `g[8]` $\sim$ `dbern(0.1111)` for $\gamma_{OHA} \sim Bernoulli(1/9)$.
- `pi` $<-$ `g[8]+0.5(1-g[8])` for $\pi = \gamma_{OHA} + 0.5(1 - \gamma_{OHA})$,
- `for (i in 5:7) { g[j]` $\sim$ `dbern(pi) }` for $\forall\ i \in \{OH, OA, HA\}$, $\gamma_j|\gamma_{OHA} \sim Bernoulli(\pi)$
- `for (j in 1:4) { g[j]` $\sim$ `dbern(1) }` for $\forall\, i \in \{, O, H, A\}$, $\gamma_j \sim Bernoulli(1)$

These priors result to prior probability for all hierarchical models equal to 1/9 and zero otherwise.

Results, using both pilot run pseudoprior and automatic pseudoprior with $k = 10$, are summarised in Table 7.1. The data give 'strong' evidence in favour of model of independence.

Model [OH][A], in which obesity and hypertension are depending on each other given the level of alcohol consumption, is the model the second highest posterior probability. All the other models have probability lower than 1%.

7.5.2 Example 2: Beetles Dataset

This dataset was used for illustration by many researchers including Spiegelhalter *et al.* (1996b). In this dataset the number of beetles killed (and the total number of beetles) after 5 hour exposure to carbon disuphlide at eight different concentrations are recorded. We consider as the response variable the number of insects killed and as explanatory the concentration. We also investigate whether the quadratic and cubic terms are significant for the model. The full model will be

$$y_i \sim Bin(n_i, p_i), \quad \eta_i = g(p_i)$$

$$\eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \quad i = 1, \ldots, 8.$$

where y_i and n_i are the number of killed insects and the total number of insects under the i exposure to carbon disuphlide; p_i denotes the probability of an insect to die after accepting i exposure to carbon disuphlide; $g(p_i)$ is the link function which is either logit, probit or complementary log-log.

In order to have similar parameter estimates and avoid high correlations between polynomial terms we use *Gram Schmidt* transformation to orthogonalize the data matrix; see, for example, Noble and Daniel (1977). Similar techniques have been applied by Clyde *et al.* (1996). The orthogonalization will accelerate the chain ensure convergence and quick mixing. The new transformed variables z_1, z_2 and z_3 are are used in the full model instead the original variables x_1, x_2 and x_3. The new 'data' matrix will be noted as $\mathbf{Z}$

As non-informative prior variance we use $\mathbf{V} = c^2 \mathbf{Z}^\mathbf{T} \mathbf{Z}$ and $c = 1.65$ according to Raftery (1996) for the logit link. The orthogonalisation of the matrix Z implies that the above priors can be independent. The prior variances for the other links are similar multiplied by an adjustment based on Taylor's expansion (for more details see Section 5).

For the logit and probit link, Gibbs variable selection strongly supports model with z_1 and z_2 in the model (about 51%). Three other models have high probabilities (from 10% to

Models	Logit (%)	Probit (%)	C.log-log (%)
$Constant$	0.00	0.00	0.00
z_1	18.43	20.19	52.66
$z_1 + z_2$	51.64	50.21	15.90
$z_1 + z_3$	12.74	13.70	25.01
$z_1 + z_2 + z_3$	17.19	15.90	6.43
Terms			
$\gamma_{z_1} = 1$	100.00	100.00	100.00
$\gamma_{z_2} = 1$	68.83	66.11	22.33
$\gamma_{z_3} = 1$	29.93	29.60	31.44

Table 7.2: Beetles Dataset: GVS Posterior Model Probabilities for each link Using BUGS (orthogonalised data, pilot-run pseudopriors, burn-in 10,000 and $10 \times 10,000$ iterations).

20%). The marginal probabilities in logit link are 100%, 68% and 30% for inclusion of z_1,z_2 and z_3 respectively. Similar are the corresponding probabilities for probit link.

The model mainly supported in complementary log-log is the model with only z_1 in the model. The main issue here is that in this link totally different terms are included in the model. The only term with probability higher that 50% is z_1.

7.5.3 Example 3: Stacks Dataset

Stacks example is a stack-loss data analysed by Spiegelhalter *et al.* (1996b) using Gibbs sampling. The dataset features 21 daily responses of stack loss (y) which measures the amount of ammonia escaping with covariates the air flow (x_1), temperature (x_2) and acid concentration (x_3). Spiegelhalter *et al.* (1996b) consider regression models with four different error structures (normal, double exponential, logistic and t_4 distributions). They also consider the cases of ridge and simple independent regression models. We extend their work by applying Gibbs variable selection on all these eight cases. The full model will be

$$y_i \sim D(\mu_i, \tau),\ \mu_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3},\ i = 1, \ldots, 21$$

where $D_i(\mu_i, \tau)$ is the distribution of the errors with mean μ_i and variance τ^{-1} which here is

	SUMMARY TABLE							
	Independence Regression				Ridge Regression			
Models	Normal	D.Exp.	Logistic	t_4	Normal	D.Exp.	Logistic	t_4
Constant	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
z_1	14.12	58.48	41.19	56.46	3.26	22.54	14.42	13.30
z_2	0.56	0.01	0.02	0.00	0.05	0.00	0.00	0.00
$z_1 + z_2$	81.25	38.64	55.25	40.46	79.79	65.00	73.32	70.92
z_3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$z_1 + z_3$	0.63	1.75	1.35	1.82	0.44	1.74	1.32	1.86
$z_2 + z_3$	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$z_1 + z_2 + z_3$	3.39	1.11	2.18	1.26	16.46	10.72	11.01	13.92
Terms								
$\gamma_{z_1} = 1$	99.30	99.98	99.97	100.00	100.00	100.00	100.00	100.00
$\gamma_{z_2} = 1$	84.90	39.76	57.45	41.72	96.50	75.72	84.33	84.84
$\gamma_{z_3} = 1$	4.30	2.86	3.53	3.08	16.10	12.46	12.33	15.78

Table 7.3: Stacks Dataset: GVS Posterior Model Probabilities Using BUGS (burn-in 10,000, samples of $10 \times 10,000$, with pilot run pseudopriors).

assumed to be normal, double exponential, logistic or t_4; where $z_{ij} = (x_{ij} - \bar{x}_j)/sd(x_j)$ are the standardised covariates. The ridge regression assumes a further restriction that the β_j for $j = 1, 2, 3$ are exchangeable (Lindley and Smith, 1972) and therefore we have $\beta_j \sim N(0, \phi^{-1})$. We use 'non-informative' priors with prior precision equal to 10^{-3} for the independent regression and for ϕ in ridge regression we use gamma prior with parameters equal to 10^{-3}. Since we do not have restrictions for the model space we use $\gamma_j \sim Bernoulli(0.5)$ for $j = 1, 2, 3$ which results to prior probability of 1/8 for all possible models. For the pilot run pseudoprior parameters we use the posterior values as given Spiegelhalter *et al.* (1996b).

Table 7.3 contains the results from all eight distinct cases using pilot run pseudopriors. In all cases flow of air (z_1) has posterior probability of inclusion higher than 99%. The temperature (z_2) seems to be also an important term with posterior probability of inclusion varying from 39% to 96%. The last term (z_3) which measures the acid concentration in air has low posterior probabilities of inclusion which are less than 5% for simple independence models and less than 20% for 'ridge' regression models.

7.5.4 Example 4: Seeds Dataset, Logistic Regression with Random Effects

This example involves the examination of a proportion of seeds that germinated on 21 plates. For these 21 plates we have recorded the seed (bean or cucumber) and the type of root extract. This data set is analysed by Spiegelhalter *et al.* (1996b) using BUGS; for more details see references there in. The model is a logistic regression with 2 categorical explanatory variables and random effects. The full model will be written

$$y_{ilk} \sim Bin(n_{ilk}, p_{ilk}), \quad log\left(\frac{p_{ilk}}{1 - p_{ilk}}\right) = m + a_i + b_l + ab_{il} + w_k, \quad i, l = 1, 2; \; k = 1, \ldots, 21.$$

where y_{ilk} and n_{ilk} is the number of seeds germinated and total number of seeds respectively for i seed l type of root extract and k plate; w_k is the random effect for the k plate.

We use sum to zero constraints for both fixed and random effects. The prior variance used here for the fixed effects is $\Sigma = 4 \times 2$. This prior is equivalent to the prior used by Dellaportas and Forster (1999) for log-linear model selection. The prior for the precision of the random effects is considered to be a gamma distribution with parameters equal to 10^{-3}. The pseudoprior parameters were taken from a pilot chain of the saturated model.

The models under consideration are ten. The prior term probabilities for the fixed effects is assigned similarly as in the example for two-way ANOVA models. For the random effects term indicator we have that $\gamma_w \sim Bernoulli(0.5)$.

	Fixed Effects		Random Effects	
Models	k=10	Pilot	k=10	Pilot
$Constant$	0.00	0.00	1.21	0.99
$[A]$	0.00	0.00	0.22	0.07
$[B]$	32.34	32.07	50.61	50.75
$[A][B]$	3.78	3.84	7.24	7.60
$[AB]$	2.80	2.83	1.80	1.85
Total	38.92	38.74	61.08	61.26

Table 7.4: Seeds Dataset: GVS Posterior Model Probabilities Using BUGS (burn-in 10,000, samples of $10 \times 10,000$).

The results in Table 7.4 give the posterior model probabilities. We used both pilot run proposals and automatic pseudoprior with $k = 10$. Both chains gave the same results as expected and the type of root extract (B) is the only factor that influences the proportion of germinated gems. The corresponding models with random and fixed effects have posterior probability equal to 51% and 32%, respectively. The marginal posterior probability of random effects is 61% which is about 56% higher than the posterior probability of fixed effects models.

7.6 Appendix of Chapter 7: BUGS CODES

Bugs code and all associated data files are freely available in electronic form at the internet web site `http://www.stat-athens.aueb.gr/~jbn/` or by electronic mail request.

7.6.1 Example 1

```
model log-linear;
#
#       3x2x4 LOG-LINEAR MODEL SELECTION WITH BUGS (GVS)
#       (c) OCTOBER 1996 ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
#       (c) REVISED OCTOBER 1997 AUEB
#
#       WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#               OF P.DELLAPORTAS AND J.J.FORSTER
#
#
const
   terms=8, # number of terms
   N = 24;  # number of Poisson cells
var
        include,    # conditional prior probability for gi
        pmdl[9],    # model indicator vector
        mdl,        # code of model
        b[N],       # model coefficients
        mean[N],    # proposal mean used in pseudoprior
        se[N],      # proposal stand.deviation used in pseudoprior
        bpriorm[N], # prior mean for b depending on g
        Tau[N,N],   # model coefficients precision
        tprior[N,N],# prior value for Tau when all terms in model
        x[N,N],     # design matrix
        z[N,N],     # matrix with z_ij=x_ij b_j g_j, used in likelihood
        n[N],       # Poisson cells
        pos[N],     # position of each parameter
        lambda[N],  # Poisson mean for each cell
        gtemp[N],   # temporary term indicator vector
        g[terms];   # term indicator vector
data pos,n in "ex2.dat", x in 'ex2des.dat', mean, se in 'prop2.dat',
tprior in 'cov.dat';
inits in "ex2.in";
{
#
#       Design Matrix can was calculated outside the bugs code
#        (e.g. via S-plus) and then can be used as data to avoid
#        useless intensive calculations. The Design matrix here was
#   calculated as a Kronecker product of stz. design matrices
#  for 3,2 and 4 levels
#
#
# associate g[i] with coefficients.
#
        for (i in 1:N) {
                gtemp[i]<-g[pos[i]];
                }
```

```
#
#       calculation of the z matrix used in likelihood
#
        for (i in 1:N) {
                for (j in 1:N) {
                        z[i,j]<-x[i,j]*b[j]*gtemp[j]
                }
        }
#
#       model configuration
        for (i in 1:N) {
                log(lambda[i])<-sum(z[i,])
                n[i]~dpois(lambda[i]);
              }
#       defining model code
#       0-> independence model [A][B][C], 1 for [AB][C], 2 for [AC][B],
#       3 for [AB][AC], 4 for [BC][A], 5 for [AB][BC], 6 for [AC][BC],
#       7 for [AB][BC], 15 for [ABC].
#
        mdl<-g[5]+2*g[6]+4*g[7]+8*g[8];
        for (i in 0:7) {
                pmdl[i+1]<-equals(mdl,i)
                }
        pmdl[9]<-equals(mdl,15)
#
#       Prior for b model coefficient
#       Mixture normal depending on current status of g[i]
#
        for (i in 1:N) { for (j in 1:N) {
#
#               GVS using se,mean from pilot run
#               *******************************
#
                Tau[i,j]<-0+tprior[i,j]*(gtemp[i]*gtemp[j])+
                          (1-gtemp[i]*gtemp[j])*equals(i,j)/(se[i]*se[i]);
#
#               Automatic proposal using prior similar to SSVS with k=10
#               ********************************************************
#                Tau[i,j]<-tprior[i,j]*pow(100,1-gtemp[i]*gtemp[j]);
#
#               Kuo and Mallick proposal is independent of g[i]
#               ********************************************
#               tau[i]=1/2   and bpriorm[i]=0
#
#                Tau[i,j]<-tprior[i,j];
#
        }
#
#               GVS PRIOR M FROM PILOT RUN
#               **************************
                bpriorm[i]<-mean[i]*(1-gtemp[i]);
#
#               PRIOR M FOR THAT DOES NOT DEPEND ON G.
#               *********************************
#                 bpriorm[i]<-0.0;
        }
        b[]~dmnorm(bpriorm[],Tau[,]);
```

```
#
#       defining prior information for gi in such way that allow
#       only hierarchical models with equal probability.
#       We also ignore models nested to the model of independent [A][B][C]
#        since we are interested in associations between factors.
#
        g[8]~dbern(0.1111111);
        include<-(1-g[8])*0.5+g[8]*1.0;
        g[7]~dbern(include);
        g[6]~dbern(include);
        g[5]~dbern(include);
        for (i in 1:4) {
                g[i]~dbern(1.0);
        }
}
```

7.6.2 Example 2

```
model beetlesgvs2;
#
#       BINOMIAL REGRESSION VARIABLE SELECTION WITH BUGS (GVS)
#          BUGS EXAMPLE: BEETLES, see BUGS examples vol.2
#           using orthogonalised data and Raftery's prior
#
#       (c) OCTOBER 1997 ATHENS UNIVERSITY OF ECONOMICS
#
#       WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#               OF P.DELLAPORTAS AND J.J.FORSTER
#
#
const
 terms=3,       # number of terms under consideration
 models=8,      # number of models
 N = 8;         # number of doses
var
 r[N],n[N],     # Binomial data, r successes out of n (total)
 p[N],          # Binomial probability of success
 x1[N],         # Data column x
 x2[N],         # Data column x^2
 x3[N],         # Data column x^3
 alpha,         # intercept coef. for original data
 beta[terms],   # model coef. for x, x^2, x^3
 t[terms],      #
 bprior[terms],# prior mean of model coef. conditional on model
 tprior[terms],# prior precision of model coef. conditional on model
 mdl,           # model index
 pmdl[models],  # model indicator
 mean[terms],   # mean of proposal from pilot run
 se[terms],     # se of proposal from pilot run
 adj,           # prior constant for link adjustment
                # based on Taylor expansion
 g[terms];      # term indicator

data r, n, x1, x2, x3 in "or.dat", mean,se in "prop4P.dat";
inits in "beetles.in";
{
 for (i in 1:N) {
```

```
   r[i] ~ dbin(p[i], n[i]);
#
#  Logit link
#  ----------
   logit(p[i])   <- alpha + g[1]*beta[1]*x1[i] + g[2]*beta[2]*x2[i]
                  + g[3]*beta[3]*x3[i];
#
#  Probit link
#  ----------
#   probit(p[i])   <- alpha + g[1]*beta[1]*x1[i] + g[2]*beta[2]*x2[i]
#                  + g[3]*beta[3]*x3[i];
#
#  Cloglog link
#  ----------
#   cloglog(p[i])   <- alpha + g[1]*beta[1]*x1[i] + g[2]*beta[2]*x2[i]
#                  + g[3]*beta[3]*x3[i];
 }
 alpha~dnorm(0,0.5)
#
# priors for glm model choice
#
#
# No adjustment is needed for logit link
  adj<-1.0
# adjustment for probit
# adj<-(4.1809*4.1809)/(2.7261*2.7261);
#
# adjustment for cloglog
  adj<-4.1809*4.1809/(2.5944*2.5944)

 t[1]<- 0.01169972*adj;
 t[2]<- 0.00003586418*adj;
 t[3]<- 9.605285E-8*adj;
 for (j in 1:terms){
#
#      ******** GVS PRIORS  ***********
#
#      GVS priors with proposals from pilot run
       bprior[j]<-(1-g[j])*mean[j];
       tprior[j] <-g[j]*t[j]+(1-g[j])*pow(se[j],-2);
#
#      GVS priors with proposals a mixture of Normals(0,c^2t^2)
#      bprior[j]<-0.0;
#      tprior[j] <-pow(100,1-g[j])*t[j];
      beta[j] ~ dnorm(bprior[j],tprior[j]); # coeffs independent
                    }
#
#       Defining Model Code
        mdl<- 1+g[1]*1+g[2]*2+g[3]*4
#
#       defining vector with model indicators
        for (j in 1:models){
        pmdl[j]<-equals(mdl,j);}
        for (i in 1:terms){g[i]~dbern(0.5)}
        }
```

7.6.3 Example 3

```
model stacks;
#
#       LINEAR REGRESSION VARIABLE SELECTION WITH BUGS (GVS)
#          BUGS EXAMPLE: STACKS, see BUGS examples vol.1
#
#       (c) OCTOBER 1997 ATHENS UNIVERSITY OF ECONOMICS
#
#       WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#               OF P.DELLAPORTAS AND J.J.FORSTER
#
#
const
   p = 3,          # number of covariates
   N = 21,         # number of observations
   models=8,       # number of models under consideration 2^8
   PI = 3.141593;
var
   x[N,p],         # raw covariates
   z[N,p] ,        # standardised covariates
   Y[N],mu[N],     # data and expectations
   stres[N],       # standardised residuals
   outlier[N],     # indicator if |stan res|  > 2.5
   beta0,beta[p], # standardised intercept, coefficients
   b0,b[p],        # unstandardised intercept, coefficients
   phi,            # prior precision of standardised coefficients
   tau,sigma,d,    # precision, sd and degrees of freedom of t distn
   g[p],           # variable indicators
   mdl,            # Model index
   pmdl[models],  # Vector with model indicators
   mean[p],se[p], # pseudoprior mean and se error
   bprior[p],      # Conditional to model Prior prior mean
   tprior[p];      # Conditional to model Prior prior precision
data Y,x in  "STACKS.DAT",
# files with proposed values
mean,se in 'pnorm.dat';   # Normal distribution
#mean,se in 'pdexp.dat';  # Double exponential distribution
#mean,se in 'plogist.dat';# Logistic distribution
#mean,se in 'pt4.dat';    # Student(4) distribution
inits in "STACKS.IN";
{
# Standardise x's and coefficients
  for (j in 1:p) {
      b[j] <- beta[j]/sd(x[,j]) ;
      for (i in 1:N) {
          z[i,j] <- (x[i,j] -  mean(x[,j]))/sd(x[,j]) ;
      }
  }
  b0 <- beta0-b[1]*mean(x[,1])-b[2]*mean(x[,2])-b[3]*mean(x[,3]);

# Model
  d <- 4;  # degrees of freedom for t
  for (i in 1:N) {
#
#     Normal Distribution
#     -------------------
```

```
      Y[i] ~ dnorm(mu[i],tau);
#
#     Double Exponential Distribution
#     -------------------
#     Y[i] ~ ddexp(mu[i],tau);
#
#     Logistic Distribution
#     -------------------
#     Y[i] ~ dlogis(mu[i],tau);
#
#     Student t4 Distribution
#     -------------------
#     Y[i] ~ dt(mu[i],tau,d);
#
      mu[i] <- beta0 + g[1]*beta[1]*z[i,1]+g[2]*beta[2]*z[i,2]
                     + g[3]*beta[3]*z[i,3];
      stres[i] <- (Y[i] - mu[i])/sigma;
#
#     if standardised residual is greater than 2.5 then outlier
      outlier[i] <- step(stres[i] - 2.5) + step(-(stres[i]+2.5) );
  }
#
#       Defining Model Code
        mdl<- 1+g[1]*1+g[2]*2+g[3]*4
#
#       defining vector with model indicators
        for (j in 1:models){
        pmdl[j]<-equals(mdl,j);}
# Priors
  beta0 ~  dnorm(0,.00001);
  for (j in 1:p) {
#
#      ******** GVS PRIORS FOR INDEPENDENCE REGRESSION ***********
#
#      GVS priors with proposals from pilot run
#       bprior[j]<-(1-g[j])*mean[j];
#       tprior[j] <-g[j]*0.001+(1-g[j])/(se[j]*se[j]);
#
#      GVS priors with proposals a mixture of Normals(0,c^2t^2)
      bprior[j]<-0.0;
      tprior[j] <-pow(100,1-g[j])*0.001;
#
#      ******** GVS PRIORS FOR RIDGE REGRESSION *****************
#
#      GVS priors with proposals from pilot run
#       bprior[j]<-(1-g[j])*mean[j];
#       tprior[j] <-g[j]*phi+(1-g[j])/(se[j]*se[j]);
#
#      GVS priors with proposals a mixture of Normals(0,c^2t^2)
#       bprior[j]<-0.0;
#       tprior[j] <-pow(100,1-g[j])*phi;
      beta[j] ~ dnorm(bprior[j],tprior[j]); # coeffs independent

  }
  tau ~ dgamma(1.0E-3,1.0E-3);
#
#  phi ~ dgamma(1.0E-3,1.0E-3);
```

```
#
#  when we use pilot run based pseudopriors bugs was unable to select
#  update method. Therefore we use an upper limit which makes bugs
#  work with Metropolis instead Gibbs
#
#  phi ~ dgamma(1.0E-3,1.0E-3)I(0,10000);
# standard deviation of error distribution
  sigma <- sqrt(1/tau);            # normal errors
# sigma <- sqrt(2)/tau;            # double exponential errors
# sigma <- sqrt(pow(PI,2)/3)/tau ;  # logistic errors
# sigma <- sqrt(d/(tau*(d-2)));     # t errors on d degrees of freedom
#
#
# Priors for variable indicators
  for (j in 1:p) { g[j]~ dbern(0.5);}
}
```

7.6.4 Example 4

```
model seedszrogvs;
#
#                LOGISTIC REGRESSION VARIABLE AND
#             RANDOM EFFECTS SELECTION WITH BUGS (GVS)
#
#           BUGS EXAMPLE: SEEDS, see BUGS examples vol.1
#
#        (c) OCTOBER 1997 ATHENS UNIVERSITY OF ECONOMICS
#
#        WRITTEN BY JOHN NTZOUFRAS UNDER THE SUPERVISION
#                OF P.DELLAPORTAS AND J.J.FORSTER
#
#

const
   terms=4,  # Number of terms under consideration
   models=16,# number of models under consideration 2^4
   N = 21;   # number of samples
var
   alpha0, alpha1, alpha2, alpha12,  # model coefficients
   tau, sigma,     # sigma= variance of random effects (tau=1/sigma)
   x1[N], x2[N],   # Design Column for factor a1 and a2
                   # here we used the STZ constraints
   p[N],           # Success probability for Binomial
   n[N],           # Total number of trials for Binomial
   r[N],           # Binomial data
   b[N],           # Random effects  (standardised)
   c[N],           # Random effects c[i] (unstandardised)
   include,        # conditional model probability for main effects
   g[terms],       # terms indicator vector
   mdl,            # model index
   pmdl[models],   # model indicator
   mean[terms-1],  # proposal mean
   se[terms-1],    # proposal se
   bprior[terms-1],# prior mean for model coefficients|model
   tprior[terms-1];# prior precision for model coefficients|model
data r,n,x1,x2 in "seeds.dat", mean,se in 'prop6.dat';
inits in "seeds.in";
```

```
{
   alpha0 ~ dnorm(0.0,1.0E-6);  # intercept

        for (j in 1:(terms-1))  {
#       ******** GVS PRIORS ***********
#
#       GVS priors with proposals from pilot run
        bprior[j]<-(1-g[j])*mean[j];
        tprior[j] <-g[j]/8+(1-g[j])/(se[j]*se[j]);
#
#       GVS priors with proposals a mixture of Normals(0,c^2t^2)
#       bprior[j]<-0.0;
#       tprior[j] <-pow(100,1-g[j])/8;
      }
#
#
   alpha1 ~  dnorm(bprior[1],tprior[1]);  # seed coeff
   alpha2 ~  dnorm(bprior[2],tprior[2]);  # extract coeff
   alpha12 ~ dnorm(bprior[3],tprior[3]);
   tau ~ dgamma(1.0E-3,1.0E-3);           # 1/sigma^2
   for (i in 1:N) {
      c[i] ~ dnorm(0.0,tau);
      b[i] <- c[i] - mean(c[]);    #  make sure b's add to zero
      logit(p[i]) <- alpha0 + g[1]*alpha1*x1[i] + g[2]*alpha2*x2[i] +
                     g[3]*alpha12*x1[i]*x2[i] + g[4]*b[i];
      r[i] ~ dbin(p[i],n[i]);
   }

   sigma <- 1.0/sqrt(tau);
#
#       Defining Model Code
        mdl<- 1+g[1]*1+g[2]*2+g[3]*4+g[4]*8
#
#       defining vector with model indicators
        for (j in 1:models){
        pmdl[j]<-equals(mdl,j);}
# Priors for variable indicators
        g[4]~ dbern(0.50);
        g[3]~ dbern(0.20);
        include<-g[3]+(1-g[3])*0.5
        g[2]~ dbern(include);
        g[1]~ dbern(include);
                    }
```

Chapter 8

Discussion and Further Research

This thesis investigated various research avenues in Bayesian model selection. Although the initial focus was on technicalities and various methodologies using MCMC, it later became evident that prior considerations are very important in Bayesian model selection and efforts strayed towards some more mathematical problems (Chapter 6).

At the time of writing, Bayesian model selection using MCMC is accepted as the leading procedure for model selection in Bayesian statistics. MCMC methods can be routinely added in standard statistical software and facilitate Bayesian model averaging techniques for safely estimating any quantities of interest. They can be used to discriminate a group of good working models and further provide a quantitative comprehensive measure of model uncertainty. The posterior probability of any model can be easily interpreted as the probability that this model is the best approximation of reality among the models considered. The flexibility to compare models that have totally different structure, the automatic notion of MCMC algorithms, the expression of posterior probabilities in simple percentages as well as the serious drawbacks of standard classical model selection techniques are the main arguments that promote the widespread of Bayesian model selection.

Future and current research is directed to the implementation of MCMC methods and especially of reversible jump samplers in specific complicated model selection issues (for example see Nobile and Green, 1997 or Vrontos *et al.* , 1998). Other general issues of MCMC need also to be investigated, such as rates of convergence (Brooks and Giudici, 1998) or automatic choices of proposals (Giudici and Roberts, 1998).

An interesting issue for further research is the discrimination between models using different quantity and quality of data. This question became evident in the models implemented in the actuarial case study presented in Chapter 1. In this case study, the first model is a simple two-way anova model using only the claim amount data with structure presented in Table 1.1. The second model incorporates the additional data of the claim count data (the number of accidents, Table 1.2). Interest lies in comparing these two models. This case study is also interesting due to the missing amounts. The construction of a MCMC algorithm that will incorporate all four models presented in this case study will enable as to directly estimate the missing counts using Bayesian model averaging techniques.

Another interesting area of future research is the use of MCMC methods presented in this thesis in the calculation of fractional and intrinsic Bayes factors. Implementation for the calculation of fractional Bayes factor seems to be straightforward and less challenging than the more computationally demanding intrinsic Bayes factor. For the latter, interest also lies in the selection of minimal samples and reference prior distributions.

Bayesian model choice is, and it will be, a very broad research area open for new mathematical and philosophical approaches. It is hoped that many of the arguments and approaches presented in this thesis may be relevant and helpful to other researchers.

Bibliography

[1] Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

[2] Aitkin, M. (1991). Posterior Bayes Factors. *Journal of the Royal Statistical Society B*, **53**, 11–142.

[3] Aitkin, M. (1995). Probability Model Choice in Single Samples from Exponential Families Using Poisson Log-linear Modelling, and Model Comparison Using Bayes and Posterior Bayes Factor. *Statistics and Computing*, **5**, 113–120.

[4] Aitkin, M. (1997). The Calibration of P-values, Posterior Bayes factors and the AIC from the Posterior Distribution of the Likelihood (with discussion). *Statistics and Computing*, **7**, 253–261.

[5] Aitkin, M., Finch, S., Mendell, N. and Thode, H. (1996). A New Test for the Presence of a Normal Mixture Distribution Based on the Posterior Bayes Factor. *Statistics and Computing*, **6**, 121–125.

[6] Akaike, H. (1969). Fitting Autoregressive Models for Control. *Annals of Mathematical Statistics*, **21**, 243–247.

[7] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*, 267–281, Budapest: Akademia Kiado.

[8] Akaike, H. (1979). A Bayesian Extension of the Minimum AIC Procedure of Autoregressive Model Fitting. *Biometrika*, **66**, 237–242.

[9] Akaike, H. (1981). Likelihood of a Model and Information Criteria. *Journal of Econometrics*, **16**, 3–14.

[10] de Alba, E., Moreno, T.M. and Juares, M. (1997). Bayesian Estimation of IBNR Reserves. *Technical Report*, I.T.A.M., Mexico.

[11] Albert, J.H. (1991). Bayesian Testing and Estimation of Association in a Two-Way Contingency Table. *Journal of the American Statistical Association*, **92**, 685–705.

[12] Albert, J.H. (1996). The Bayesian Selection of Log-linear Models. *Canadian Journal of Statistics*, **24**, 327–347.

[13] Albert, J.H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669–679.

[14] Albert, J.H. and Chib, S. (1997). Bayesian Tests and Model Diagnostics in Conditionally Independent Hierarchical Models. *Journal of the American Statistical Association*, **92**, 916–925.

[15] Atkinson, A.C. (1978). Posterior Probabilities for Choosing a Regression Model. *Biometrika*, **65**, 39–48.

[16] Atkinson, A.C. (1980). A Note on the Generalized Information Criterion for Choice of a Model. *Biometrika*, **67**, 413–418.

[17] Atkinson, A.C. (1981). Likelihood Ratios, Posterior Odds and Information Criteria. *Journal of Econometrics*, **16**, 15–20.

[18] Barnett, G., Kohn, R. and Sheather, S. (1996). Bayesian Estimation of an Autoregressive Model Using Markov Chain Monte Carlo. *Journal of Econometrics*, **74**, 237–254.

[19] Bartlett, M.S. (1957). Comment on D.V. Lindley's Statistical Paradox. *Biometrika*, **44**, 533–534.

[20] Basu and Mukhopadhyay (1994). Bayesian Analysis of a Random Link Function in Binary Response Regression. *Technical Report*, Department of Mathematical Sciences, University of Arkansas, USA.

[21] Bayarri, S. and Berger, J.O. (1998a). Measures of Surprise in Bayesian Analysis. *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[22] Bayarri, S. and Berger, J.O. (1998b). Quantifying Surprise in the Data and Model Verification. *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[23] Bayarri, S. and Berger, J.O. (1998c). Quantifying Surprise in the Data. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[24] Bendel, R.B. and Afifi, A.A. (1977). Comparison of Stopping Rules in Forward "Stepwise" Regression. *Journal of the American Statistical Association*, **72**, 47–53.

[25] Bedrick, E.J., Christensen, R. and Johnson, W. (1996). A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association*, **91**, 1450–1460.

[26] Berger, J.O., Brown, L.D. and Wolpert, R.L. (1994). A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing. *Annals of Statistics*, **22**, 1787–1807.

[27] Berger, J.O., Bookai, B. and Wang, Y. (1997). Unified Frequentist and Bayesian Testing of a Precise Hypothesis. *Statistical Science*, **12**, 133–160.

[28] Berger, J.O. and Delampady, M. (1987). Testing Precise Hypotheses. *Statistical Science*, **2**, 317–352.

[29] Berger, J.O. and Montera, J. (1991). Interpreting the Stars in Precise Hypothesis Testing. *International Statistical Review*, **59**, 337–353.

[30] Berger, J.O. and Montera, J. (1998). Default Bayes Factors for One-Sided Hypothesis Testing. *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[31] Berger, J.O. and Pericchi, L.R. (1996a). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, **91**, 109–122.

[32] Berger, J.O. and Pericchi, L.R. (1996b). The Intrinsic Bayes Factor for Linear Models. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 25–44.

[33] Berger, J.O. and Pericchi, L.R. (1998). On Criticisms and Comparisons of Default Bayes Factors for Model Selection and Hypothesis Testing. *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[34] Berger, J.O. and Pericchi, L.R. (1999). Accurate and Stable Bayesian Model Selection: The Median Intrinsic Bayes Factor. *Sankhyā B*, **60**, 1–18.

[35] Berger, J.O. and Selke, T. (1987). Testing a Point Null Hypothesis: Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, **82**, 112–139.

[36] Bernardo, J.M. (1997). Non-informative Priors Do not Exist: A Dialog with Jose M. Bernardo (Interviewers: T.Z. Irony and N.D. Singpurwalla). *Journal of Statistical Planning and Inference*, **65**, 159–189.

[37] Bernardo, J.M. (1999). Nested Hypothesis Testing: the Bayesian Reference Criterion. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[38] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Chichester: John Wiley & Sons.

[39] Best, N., Cowles, M.K. and Vines, K. (1995). *CODA: Convergence Diagnostics and Output Analysis Software for Gibbs Sampling Output.* MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

[40] Bhansali, R.J. (1997). Direct Autoregressive Predictors for Multistep Prediction: Order Selection and Performance Relative to the Plug in Predictors. *Statistica Sinica*, **7**, 425–449.

[41] Bhansali, R.J. and Downham, D.Y. (1977). Some Properties of the order of an Autoregression Model Selected by a Generalization of Akaike's EPF Criterion. *Biometrika*, **64**, 547–551.

[42] Box, G.E.P. (1980). Sampling and Bayes Inference in Scientific Modeling and Robustness. *Journal of the Royal Statistical Society A*, **143**, 380–430.

[43] Brewer, K.R.W. (1998). Reference Bayesian Hypothesis Testing. *Technical Report*, Australian National University, Australia.

[44] Brooks, S.P. and Giudici, P. (1998). Convergence Assessment for Reversible Jump MCMC Simulations. *Technical Report*, University of Bristol, UK.

[45] Brooks, S.P. and Roberts, G.O. (1997). Assessing Convergence of Markov Chain Monte Carlo. *Technical Report*, University of Bristol, UK.

[46] Brown, P.J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian Variable Selection and Prediction. *Journal of the Royal Statistical Society B*, **60**, 627–641.

[47] Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model Selection: An Integral Part of Inference. *Biometrics*, **53**, 603–618.

[48] Carlin, B.P. (1992). State Space Modeling of Non-standard Actuarial Time Series. *Insurance: Mathematics and Economics*, **11**, 209–222.

[49] Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society B*, **157**, 473–484.

[50] Carlin, B.P., Polson, N.G. and Stoffer, D.S. (1992). A Monte Carlo Approach to Nonnormal and nonlinear State-Space Modeling. *Journal of the American Statistical Association*, **87**, 493–500.

[51] Carter, C.K. and Kohn, R. (1994). On Gibbs Sampler for State Space Models. *Biometrika*, **81**, 541–553.

[52] Casella, G. and Berger, R.L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*, **82**, 106–111.

[53] Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, **46**, 167–174.

[54] Cavanaugh, J.E. and Shumway, R.H. (1997). A Bootstrap Variant of AIC for State-Space Model Selection. *Statistica Sinica*, **7**, 473–496.

[55] Chatfield, C. (1995) Model Uncertainty, Data Mining and Statistical Inference (with discussion). *Journal of the Royal Statistical Society A*, **158**, 419–466.

[56] Chen, M.H., Ibrahim, J.G. and Yiannoutsos, C. (1999). Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models. *Journal of the Royal Statistical Society B*, **61**, 223–243

[57] Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313–1321.

[58] Chib, S. and Greenberg, E. (1995), Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, **49**, 327–335.

[59] Chipman, H. (1996). Bayesian Variable Selection with Related Predictors. *Canadian Journal of Statistics*, **24**, 17–36.

[60] Chipman, H. (1997). Fast Model Search for Designed Experiments with Complex Aliasing. *Technical Report*, Graduate School of Business, University of Chicago, USA.

[61] Chipman, H., Hamada, M. and Wu, C.F.J. (1997). A Bayesian Variable Selection Approach for Analysing Designed Experiments with Complex Aliasing. *Technometrics*, **39**, 372–381.

[62] Chow, G.C. (1981). A Comparison of the Information and Posterior Probability Criteria for Model Selection. *Journal of Econometrics*, **16**, 21–33.

[63] Clyde, M.A. (1999). Bayesian Model Averaging and Model Search Strategies. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[64] Clyde, M. and DeSimone-Sasinowska, H. (1997). Accounting for Model Uncertainty in Poisson Regression Models: Does Particulate Matter? *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[65] Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via Orthogonalized Model Mixing. *Journal of the American Statistical Association*, **91**, 1197–1208.

[66] Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple Shrinkage and Subset Selection in Wavelets. *Biometrika*, **85**, 391-402.

[67] Conigliani, C. and O'Hagan, A. (1996). Sensitivity Measures of the Fractional Bayes Factor. *Technical Report*, University of Nottingham, UK.

[68] Copas, J.B. (1984). Discussion of Dr Miller's Paper. *Journal of the Royal Statistical Society A*, **,** 147, 410-412.

[69] Cowles, M.K. and Carlin, B.P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, **91**, 883-904.

[70] De Jong, P. and Zehnwirth, B. (1983). Claims Reserving, State Space Models and Kalman Filter. *The Journal of the Institute of Actuaries*, **110**, 157–181.

[71] Delampady, M. and Berger, J.O. (1990). Lower Bounds on Bayes Factors for Multinomial Distributions with Application to Chi-squared Tests of Fit. *Annals of Statistics*, **18**, 1295–1316.

[72] Dellaportas, P. and Forster, J.J. (1999). Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-linear Models. *Biometrika*, (to appear).

[73] Dellaportas, P., Forster, J.J. and Ntzoufras, I. (1998). On Bayesian Model and Variable Selection Using MCMC. *Technical Report* **40**, Department of Statistics, Athens University of Economics and Business, Greece. (submitted).

[74] Dellaportas, P., Forster, J.J. and Ntzoufras, I. (1999). Bayesian Variable Selection Using the Gibbs Sampler. *Generalized Linear Models: A Bayesian Perspective* (D. K. Dey, S. Ghosh, and B. Mallick, eds.). New York: Marcel Dekker, (to appear).

[75] Dellaportas, P., Karlis, D. and Xekalaki, E. (1998). Bayesian Analysis of Finite Poisson Mixtures. *Technical Report* , Department of Statistics, Athens University of Economics and Business, Greece.

[76] Dellaportas, P. and Smith, A.F.M. (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Applied Statistics*, **42**, 443–460.

[77] Dempster, A.P. (1974). The Direct Use of Likelihood for Significance Testing. *Proc. Conf. Foundational Questions in Statistical Inference* (O. Barndorff-Nielsen, P. Blaesild and G. Sihon, eds.). University of Aarhus, 335–352. [Reprinted: 1997, *Statistics and Computing*, **7**, 247–252].

[78] Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998a). A Bayesian CART Algorithm. *Biometrika*, **85**, 363–377.

[79] Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998b). Bayesian MARS. *Technical Report*, Department of Mathematics, Imperial College of Science, Technology and Science, London.

[80] De Santis, F. and Spezzaferri, F. (1997). Methods for Default and Robust Bayesian Model Comparison: The Fractional Bayes Factor Approach. *Technical Report*, Department of Statistics, Carnegie Mellon University, USA.

[81] DiCiccio, T.J., Kass, R.E., Raftery, A. and Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, **92**, 903–915.

[82] Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with discussion). *Journal of the Royal Statistical Society B*, **57**, 45–97.

[83] Draper, N.R. and Smith, H. (1966). *Applied Regression Analysis.* New York: John Wiley & Sons.

[84] Dudley, R.M. and Haughton, D. (1997). Information Criteria for Multiple Datasets. *Statistica Sinica*, **7**, 265–284.

[85] Edwards, D. and Havránek, T. (1985). A Fast Procedure for Model Search in Multidimensional Contingency Tables. *Biometrika*, **72**, 339–351.

[86] Efroymson, M.A. (1960). Multiple Regression Analysis. *Mathematical Methods for Digital Computers*, Vol. 1 (A. Ralston and H.S. Wilf, eds.). New York: John Wiley & Sons, 191–203.

[87] Efroymson, M.A. (1966). Stepwise Regression - a Backward and Forward Look. *Presented at the Easter Regional Meetings of the Inst. of Meth.Statist.*, Florham Park, New Jersey.

[88] Erkanli, A. (1994) Laplace Approximations for Posterior Expectation When the Model Occurs at the Boundary of the Parameter Space. *Journal of the American Statistical Association*, **89**, 250–258.

[89] Evans, M. (1997). Bayesian Inference Procedures Derived via the Concept of Relative Surprise. *Communication in Statistics: Theory and Methods*, **26**, 1125–1143.

[90] Evans, M. and Swartz, T.B. (1996). Discussion of Methods for approximating integrals in Statistics with special emphasis on Bayesian integration problems. *Statistical Science*, **11**, 54–64.

[91] Fernandez, C., Ley, E. and Steel, M.F.J. (1998). Benchmark Priors For Bayesian Model Averaging. *Technical Report*, Department of Econometrics, Tilburg University, the Netherlands.

[92] Fernandez, C., Steel, M.F.J. and Ley, E. (1997). Statistical Modeling of Fishing Activities in the North Atlantic. *Technical Report*, Department of Econometrics, Tilburg University, the Netherlands.

[93] Foster, D.P. and George, E.I. (1994). The Risk Inflation Criterion for Multiple Regression. *Annals of Statistics*, **22**, 1947–1975.

[94] Freedman, D. (1983). A Note on Screening Regression Equations. *The American Statistician*, **37**, 152–155.

[95] Gamerman, D. (1998). Markov Chain Monte Carlo for Dynamic Generalized Linear Models. *Biometrika*, **85**, 215–227.

[96] Garthwaite, P.H. and Dickey, J.M. (1992). Elicitation of Prior Distributions for Variable-Selection Problems in Regression. *Annals of Statistics*, **20**, 1697–1719.

[97] Geiger, D., Heckerman, D. and Meek, C. (1996). Asymptotic Model Selection for Directed Networks with Hidden Variables. *Technical Report* **96-07**, Microsoft Research, Advanced Technology Division, Microsoft Corporation, USA.

[98] Gelfand, A.E. and Dey, D.K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society B*, **56**, 501–514.

[99] Gelfand, A.E. and Ghosh, S.K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, **85**, 1–13.

[100] Gelfand, A.E., Hills, S.E., Racine-Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian Inference in Normal Data Models using Gibbs Sampling. *Journal of the American Statistical Association*, **85**, 972–985.

[101] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85**, 398–409.

[102] Gelfand, A.E., Smith, A.F.M. and Lee, T-M. (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, **87**, 523–532.

[103] Gelman, A. and Meng, X.L. (1996). Model Checking and Model Improvement. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 189–202.

[104] Gelman, A., Meng, X.L. and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, **6**, 733–807.

[105] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Trans. Information Theory*, **6**, 721–741.

[106] George, E.I. and Foster, D.P. (1997). Calibration and Empirical Bayes Variable Selection. *Technical Report*, University of Texas at Austin and University of Pennsylvania, USA.

[107] George, E.I. and McCulloch, R.E. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.

[108] George, E.I. and McCulloch, R.E. (1996). Stochastic Search Variable Selection. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 203–214.

[109] George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, **7**, 339–373.

[110] George, E.I., McCulloch, R.E. and Tsay, R.S. (1996). Two Approaches to Bayesian Model Selection with Applications. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (D. A.Berry, K. M. Chaloner and J. K. Geweke, eds.). New York: John Wiley & Sons, 339–348.

[111] Geweke, J. (1996). Variable Selection and Model Comparison in Regression. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 609–620.

[112] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J (1996). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

[113] Gilks, W.R. and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics*, **41**, 337–348.

[114] Giudici, P. and Green, P.J. (1998). Decomposable Graphical Gaussian Model Determination. *Technical Report*, University of Bristol, UK.

[115] Giudici, P. and Roberts, G. (1998). On the Automatic Choice of Reversible Jumps. *Technical Report*, University of Pavia, Italy.

[116] Godsill, S. (1998). On the Relationship Between MCMC Model Uncertainty Methods. *Technical Report* **305**, Signal Processing and Communications Laboratoty, Department of Engineering, University of Cambridge, UK.

[117] Gorman, J.W. and Toman, R.J. (1966). Selection of Variables for Fitting Equation to Data. *Technometrics*, **8**, 27–51.

[118] Goutis, C. and Robert, C.P. (1998). Model Choice in Generalized Linear Models: A Bayesian Approach via Kullback-Leibler Projections. *Biometrika*, **82**, 711–732.

[119] Green, P. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, 711–732.

[120] Guttman, I. (1967). The Use of the Concept of a Future Observation in Goodness-of-fit Problems. *Journal of the Royal Statistical Society B*, **29**, 83–100.

[121] Haastrup, S. and Arjas, E. (1996). Claims Reserving in Continuous time; A Nonparametric Bayesian Approach. *Astin Bulletin*, **26**, 139–164.

[122] Haberman, S. and Renshaw, A.E. (1996). Generalized Linear Models and Actuarial Science. *The Statistician*, **45**, 407–436.

[123] Hannan, E.J. and Quinn, B.G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society B*, **41**, 190–195.

[124] Hastings, W.K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**, 97–109.

[125] Healy, M.J.R. (1988). *Glim: An Introduction.* Oxford: University Press.

[126] Heckerman, D. and Meek, C. (1997). Models and Selection Criteria for Regression and Classification. *Technical Report* **97-08**, Microsoft Research, Advanced Technology Division, Microsoft Corporation, USA.

[127] Hesselager, O. (1991). Prediction of Outstanding Claims: An Hierarchical Credibility Approach. *Scandinavian Actuarial Journal*, **1991**, 25–47.

[128] Hocking, R.R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, **32**, 1–51.

[129] Hoeting, J.A. and Ibrahim, J.G (1997). Bayesian Predictive Simultaneously Variable and Transformation Selection in the Linear Model. *Technical Report*, Department of Statistics, Colorado State University, USA.

[130] Hoeting, J.A., Madigan, D. and Raftery, A.E. (1995). Simultaneously Variable and Transformation Selection in Linear Regression. *Technical Report* , Department of Statistics, University of Washington, USA.

[131] Hoeting, J.A., Madigan, D. and Raftery, A.E. (1996). A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression. *Journal of Computational Statistics and Data Analysis*, **22**, 251-270.

[132] Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1998). Bayesian Model Averaging (review paper). *Technical Report* **335**, Department of Statistics, University of Washington, USA.

[133] Hurvich, C.M. and Tsai, C.L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, **76**, 297–307.

[134] Hurvich, C.M. and Tsai, C.L. (1991). Bias of the Corrected AIC Criterion for Underfitted Regression and Time Series Models. *Biometrika*, **78**, 499–509.

[135] Ibrahim, J.G. and Chen, M.H. (1997). Predictive Variable Selection for the Multivariate Linear Model. *Biometrics*, **53**, 465–478.

[136] Ibrahim, J.G. and Chen, M.H. (1999). Prior Elicitation and Variable Selection for Generalized Mixed Models. *Generalized Linear Models: A Bayesian Perspective* (D. K. Dey, S. Ghosh, and B. Mallick, eds.). New York: Marcel Dekker, (to appear).

[137] Ibrahim, J.G. and Laud, P.W. (1994). A Predictive Approach to the Analysis of Designed Experiments. *Journal of the Royal Statistical Society B*, **89**, 309–319.

[138] Ibrahim, J.G. and Laud, P.W. (1996). Predictive Approaches to Bayesian Model Selection. *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner* (D. A.Berry, K. M. Chaloner and J. K. Geweke, eds.). New York: John Wiley & Sons, 349–358.

[139] Jefferys, W.H. and Berger, J.O. (1991). Sharpening Ockham's Razor on Bayesian Strop. *Technical Report* **91-44**, Department of Statistics, Purdue University, USA.

[140] Jewell, W.S. (1989). Predicting IBNYR Events and Delays. *Astin Bulletin*, **19**, 25–55.

[141] Kass, R.E. (1993). Bayes Factors in Practice. *The Statistician*, **42**, 551–560.

[142] Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, **90**, 773–795.

[143] Kass, R.E. and Wasserman, L. (1995). A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, **90**, 928–934.

[144] Kass, R.E. and Wasserman, L. (1996). The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

[145] Kennedy, W.J. and Bancroft, T.A. (1975). Model-Building for Prediction in Regression Based on Repeated Significance Tests. *Annals of Mathematical Statistics*, **42**, 1273–1284.

[146] Key, J.T. (1996). Studies of a Simulation Approach to Bayesian Model Comparison. *Ph.D. Thesis*, Department of Mathematics, Imperial College of Science, Technology and Science, London.

[147] Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1997). Choosing among Models When None of them Are True. *Technical Report*, Department of Mathematics, Imperial College of Science, Technology and Science, London.

[148] Key, J.T., Pericchi, L.R. and Smith, A.F.M. (1998). Bayesian Model Choice: What and Why? *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[149] Knuiman, M.W. and Speed, T.P. (1988). Incorporating Prior Information Into the Analysis of Contingency Tables. *Biometrics*, **44**, 1061–1071.

[150] Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhyā B*, **60**, 65–81.

[151] Lai, T.L. and Lee, C.P. (1997). Information and Prediction Criteria for Model Selection in Stochastic Regression and ARMA Models. *Statistica Sinica*, **7**, 285–309.

[152] Lang, J.B. (1997). Bayesian Ordinal and Binary Regression Models with a Parametric Family of Mixture Links. *Technical Report* **267**, Department of Statistics and Actuarial Science, University of Iowa, USA.

[153] Laud, P.W. and Ibrahim, J.G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society B*, **57**, 247–262.

[154] Laud, P.W. and Ibrahim, J.G. (1996). Predictive Specification of Prior Model Probabilities in Variable Selection. *Biometrika*, **83**, 267–274.

[155] Lawless, J.F. (1994). Adjustments for Reporting Delays and the Prediction of Occurred but not Reported Events. *Canadian Journal of Statistics*, **22**, 15-31.

[156] Leamer, E.E. (1978). *Specification Searches.* New York: John Wiley & Sons.

[157] Lewis, S.M. and Raftery, A.E. (1996). Comment: Posterior Predictive Assessment for Data Subsets in Hierarchical Models via MCMC. *Statistica Sinica*, **6**, 733–807.

[158] Lindley, D.V. (1957). A Statistical Paradox. *Biometrika*, **44**, 187–192.

[159] Lindley, D.V. (1968). The Choice of Variables in Regression. *Journal of the Royal Statistical Society B*, **30**, 31–66.

[160] Lindley, D.V. (1980). L.J.Savage - His Work in Probability and Statistics. *Annals of Statistics*, **8**, 1–24.

[161] Lindley, D.V. (1993). On Presentation of Evidence. *Mathematical Scientist*, **18**, 60–63.

[162] Lindley, D.V. and Smith, A.F.M. (1972). Bayes Estimates for the Linear Model (with discussion). *Journal of the Royal Statistical Society B*, **34**, 1–41.

[163] Liu, J.S. (1996a). Metropolised Gibbs Sampler: An Improvement. *Technical Report*, Department of Statistics, Stanford University, California, USA.

[164] Liu, J.S. (1996b). Peskun's Theorem and a Modified Discrete-State Gibbs Sampler. *Biometrika*, **83**, 681–682.

[165] Madigan, D. and Raftery, A.E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window. *Journal of the American Statistical Association*, **89**, 1535–1546.

[166] Madigan, D., Raftery, A.E. Volinsky, C.T. and Hoeting, J.A. (1996). Bayesian Model Averaging. *AAAI Workshop on Integrating Multiple Learned Models*, **1996**, 77-83

[167] Madigan, D., Raftery, A.E., York, J., Bradshaw, J.M. and Almond, R.G. (1995). Strategies for Graphical Model Selection. *Selecting Models from Data: AI and Statistics IV* (P. Cheesman and R. W. Oldford, eds.). Berlin: Springer Verlag, 91–100.

[168] Madigan, D. and York, J. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review*, **63**, 215–232.

[169] Makov, U.E., Smith, A.F.M. and Liu, Y.H. (1996). Bayesian Methods in Actuarial Science. *The Statistician*, **45**, 503–515.

[170] Mallick, B.K. and Gelfand, A.E. (1994). Generalized Linear Models with Unknown Number of Components. *Biometrika*, **81**, 237–245.

[171] Mallows, C.L. (1973). Some Comments on C_p. *Technometrics*, **15**, 661–676.

[172] McCullagh, P. and Nelder, J.A. (1983). Generalized Linear Models. London: Chapman and Hall.

[173] Meng, X.L. (1994). Posterior Predictive P-Values. *Annals of Statistics*, **22**, 1142–1160.

[174] Metropolis, N., Rosenbluth, A., Rosenbluth, M.N., Teller A.H. and Teller E. (1953). Equations of State Calculations by Fast Computing Machine. *J. Chem. Phys.*, **21**, 1087–1091.

[175] Miller, A.J. (1984). Selection of Subsets of Regression Variables. *Journal of the Royal Statistical Society A*, **147**, 389–425.

[176] Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, **83**, 1023–1036.

[177] Neuhaus, W. (1992). IBNR Models with Random Delay Distributions. *Scandinavian Actuarial Journal*, **1992**, 97–107.

[178] Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society B*, **56**, 3–48.

[179] Nobile, A. and Green, P.J. (1997). Bayesian Analysis of Factorial Experiments by Mixture Modelling. *Technical Report*, University of Bristol, UK.

[180] Noble, B. and Daniel, J.W. (1977). *Applied Linear Algevra.* Englewood Cliffs, New Jersey: Prentice-Hall.

[181] Norberg, R. (1986). A Contribution to Modelling of IBNR Claims. *Scandinavian Actuarial Journal*, **1986**, 155–203.

[182] Ntzoufras, I. (1995). Model Choice Methods for High Dimensional Contingency Tables. *M.Sc. Dissertation*, Department of Mathematics, University of Southampton, UK.

[183] Ntzoufras, I. (1999a). Discussion on 'Bayesian Model Averaging and Model Search Strategies'. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, (to appear).

[184] Ntzoufras, I. (1999b). Gibbs Variable Selection Using BUGS. *Technical Report* , Department of Statistics, Athens University of Economics and Business, Greece.

[185] Ntzoufras, I. and Dellaportas, P. (1998). Bayesian Modelling of Outstanding Liabilities Incorporating Claim Count Uncertainty. *Technical Report* **41**, Department of Statistics, Athens University of Economics and Business, Greece. (submitted).

[186] Ntzoufras, I., Dellaportas, P. and Forster, J.J. (1996). A Comparison of Markov Chain Monte Carlo Methods for Log-Linear Model Selection. *Proceedings of the Third Hellenic-European Conference on Mathematics and Informatics* (E.A.Lipitakis, eds.). Athens: LEA, 506–514.

[187] Ntzoufras, I., Dellaportas, P. and Forster, J.J. (1999a). Specification and Interpretation of Prior Distributions for Variable Selection in Linear Models. *Proceedings of the*

Fourth Hellenic-European Conference on Computer Mathematics and its Applications (E.A.Lipitakis, eds.). Athens: LEA, (to appear).

[188] Ntzoufras, I., Dellaportas, P. and Forster, J.J. (1999b). MCMC Variable and Link Determination in Generalised Linear Models. *Technical Report* , Department of Statistics, Athens University of Economics and Business, Greece.

[189] Ntzoufras, I., Forster, J.J. and Dellaportas, P. (1998). Stochastic Search Variable Selection for Log-linear Models. *Technical Report* **38**, Department of Statistics, Athens University of Economics and Business, Greece. (submitted).

[190] O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics vol.2b: Bayesian Inference*. London: Edward Arnold.

[191] O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society B*, **57**, 99–138.

[192] O'Hagan, A. (1997). Properties of Intrinsic and Fractional Bayes Factors. *Test*, **6**, 101–118.

[193] Panaretos, J., Psarakis, S. and Xekalaki, E. (1997). The Correlated Gamma-Ratio Distribution in Model Evaluation and Selection. *Technical Report* **33**, Department of Statistics, Athens University of Economics and Business, Greece.

[194] Pauler, D.K. (1998). The Schwarz Criterion and Related Methods for Normal Linear Models. *Biometrika*, **85**, 13–27.

[195] Pauler, D.K., Wakefield, J.C. and Kass, R.E.(1998). Bayes Factors for Variance Components Models. *Technical Report*, Department of Statistical Science, University College London, UK.

[196] Petris, G. and Tardella, L. (1998). Simulating from Mixture Distributions with Applications to Bayesian Model Selection. *Technical Report*, Department of Statistics, Carnegie Mellon University, USA.

[197] Pericchi, L.R. (1984). An Alternative to the Standard Bayesian Procedure for Discrimination Between Normal Linear Models. *Biometrika*, **71**, 575–586.

[198] Poskitt, D.S. (1987). Precision, Complexity and Bayesian Model Determination. *Journal of the Royal Statistical Society B*, **49**, 199–208.

[199] Poskitt, D.S. and Tremayne, A.R. (1983). On the Posterior Odds of Time Series Models. *Biometrika*, **70**, 157–162.

[200] Raftery, A.E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology 1995* (P. V. Marsden ed.). Oxford: Blackwell.

[201] Raftery, A.E. (1996a). Approximate Bayes Factors and Accounting for Model Uncertainty in Generalized Linear Models. *Biometrika*, **83**, 251–266.

[202] Raftery, A.E. (1996b). Hypothesis Testing and Model Selection. *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.). London: Chapman and Hall, 163–188.

[203] Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, **92**, 179–191.

[204] Raftery, A.E., Madigan, D., and Volinsky, C.T. (1996). Accounting for Model Uncertainty in Survival Analysis Improves Predictive Performance. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 323–349.

[205] Raftery, A.E. and Richardson, S. (1996). Model Selection for Generalized Linear Models via GLIB: Application to Nutrition and Breast Cancer. *Bayesian Biostatistics* (D.A. Berry and D.K. Strangl, eds.). New York: Marcel Dekker, 321–353.

[206] Rao, C.R. and Wu, Y. (1989). A Strongly Consistent Procedure for Model Selection in a Regression Problem. *Biometrika*, **76**, 369–374.

[207] Renshaw, A.E. (1989). Chain Ladder and Interactive Modelling. *The Journal of the Institute of Actuaries*, **116**, 559–587.

[208] Renshaw, A.E. (1994). On the Second Moment Properties and the Implementation of Certain GLIM Based Stochastic Claims Reserving Models. *Technical Report* **65**, Department of Actuarial and Statistics, City University, London, UK.

[209] Renshaw, A.E. and Verrall, R. (1994). A Stochastic Model Underlying the Chain-Ladder Technique. *Proceedings XXV ASTIN Colloquium*, Cannes, France.

[210] Richardson, S. and Green, P.J. (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society B*, **59**, 731–792.

[211] Rissanen, J. (1986). Order Estimation by Accumulated Prediction Errors. *Journal of Applied Probability*, **23**, 55-61.

[212] Robert, C.P. (1993). A Note on Jeffreys-Lindley Paradox. *Statistica Sinica*, **3**, 601-608.

[213] Ronchetti, E. (1997). Robustness Aspects of Model Choice. *Statistica Sinica*, **7**, 327–338.

[214] Rubin, D.B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics*, **12**, 1152–1172.

[215] San Martini, A. and Spezzaferri, F. (1984). A Predictive Model Selection Criterion. *Journal of the Royal Statistical Society B*, **46**, 296–303.

[216] Schervish, M.J. (1996). P-values: What They Are and What They Are Not. *The American Statistician*, **50**, 203–206.

[217] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.

[218] Shafer, J. (1982). Lindley's Paradox (with discussion). *Journal of the American Statistical Association*, **77**, 325–334.

[219] Shao, J. (1993). Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, **88**, 486–494.

[220] Shao, J. (1997). An Asymptotically Theory for Linear Model Selection (with discussion). *Statistica Sinica*, **7**, 221–264.

[221] Shi, P. and Tsai, C.L. (1998). A Note on the Unification of the Akaike Information Criterion. *Journal of the Royal Statistical Society B*, **60**, 551–558.

[222] Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. *Annals of Statistics*, **8**, 147–164.

[223] Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika*, **68**, 45–54.

[224] Shibata, R. (1984). Approximate Efficiency of a Selection Procedure for the Number of Regression Variables. *Biometrika*, **71**, 43–49.

[225] Shibata, R. (1997). Bootstrap Estimate of the Kullback-Leibler Information for Model Selection. *Statistica Sinica*, **7**, 375–394.

[226] Smith, A.F.M. and Spiegelhalter, D.J. (1980). Bayes Factor and Choice Criteria for the Linear Models. *Journal of the Royal Statistical Society B*, **42**, 213–220.

[227] Smith, A.F.M. and Roberts, G.O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society B*, **55**, 3–23.

[228] Smith, M. and Kohn, R. (1996). Nonparametric Regression Using Bayesian Variable Selection. *Journal of Econometrics*, **75**, 317–343.

[229] Smith, M. and Kohn, R. (1997). A Bayesian Approach to Nonparametric Bivariate Regression. *Journal of the American Statistical Association*, **92**, 1522–1535.

[230] Smith, M., Wong, C.M. and Kohn, R. (1998). Additive Nonparametric Regression with Autocorrelated Errors. *Journal of the Royal Statistical Society B*, **60**, 311–331.

[231] Spiegelhalter, D.J. and Smith, A.F.M. (1982). Bayes Factors for Linear and Log-linear Models with Vague Prior Information. *Journal of the Royal Statistical Society B*, **44**, 377-387.

[232] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996a). *BUGS 0.5: Bayesian Inference Using Gibbs Sampling Manual.* MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

[233] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996b). *BUGS 0.5: Examples Volume 1.* MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

[234] Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W.(1996c). *BUGS 0.5: Examples Volume 2.* MRC Biostatistics Unit, Institute of Public health, Cambridge, UK.

[235] Stangl, D.K. (1996). Bayesian Methods in the Analysis of Clinical Trials: A Discussion. *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[236] Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society B*, **39**, 44-47.

[237] Taylor, G.C. and Ashe, F.R. (1983). Second Moments of estimates of Outstanding Claims. *Journal of Econometrics*, **23**, 37–61.

[238] Thompson, M.L. (1978). Selection of Variables in Multiple Regression: Part I. A Review and Evaluation. *International Statistical Review*, **46**, 1–19.

[239] Tierney, L. and Kadane, J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, **81**, 82–86.

[240] Tierney, L., Kass, R.E. and Kadane, J.B. (1989). Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of the American Statistical Association*, **84**, 710–716.

[241] Troughton, P.T. and Godsill, S.J. (1997). A Reversible Jump Sampler for Autoregressive Time Series, Employing Full Conditionals to Achieve Efficient Model Space Moves. *Technical Report* **304**, Signal Processing and Communications Laboratoty, Department of Engineering, University of Cambridge, UK.

[242] Varshavsky, J.A. (1996). Intrinsic Bayes Factors for Model Selection with Autoregressive Data. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford: University Press, 757–764.

[243] Verdinelli, I. and Wasserman, L. (1995). Computing Bayes Factors Using a Generalization of the Savage Dickey Density Ratio. *Journal of the American Statistical Association*, **90**, 614–618.

[244] Verrall, R. (1989). State Space Representation of the Chain Ladder Linear Model. *The Journal of the Institute of Actuaries*, **116**, 589–610.

[245] Verrall, R. (1990). Bayes and Empirical Bayes Estimation for the Chain Ladder Model. *Astin Bulletin*, **20**, 217–243.

[246] Verrall, R. (1991). Chain Ladder and Maximum Likelihood. *The Journal of the Institute of Actuaries*, **118**, 489–499.

[247] Verrall, R. (1993). Negative Incremental Claims: Chain Ladder and Linear Models. *The Journal of the Institute of Actuaries*, **120**, 171–183.

[248] Verrall, R. (1994). A Method for Modelling Varying-off Evolutions in Claims Reserving. *Astin Bulletin*, **24**, 325–332.

[249] Verrall, R. (1996). Claims Reserving and Generalized Additive Models. *Insurance: - Mathematics and Economics*, **19**, 31–43.

[250] Volinsky, C., Madigan, D., Raftery, A.E. and Kronmal, R. (1996). Bayesian Model Averaging in Proportional Hazard models: Assessing Stroke Risk. *Technical Report* **302**, Department of Statistics, University of Washington, USA.

[251] Volinsky, C., Madigan, D., Raftery, A.E. and Kronmal, R. (1997). Bayesian Model Averaging in Proportional Hazard models: Assessing Stroke Risk. *Applied Statistics*, **46**, 443–448.

[252] Vrontos, I.D., Dellaportas, P. and Politis, D.N. (1998). Full Bayesian Inference for GARCH and EGARCH Models. *Technical Report* **37**, Department of Statistics, Athens University of Economics and Business, Greece..

[253] Wakefield, J. and Bennett, J. (1996). The Bayesian Modelling of Covariates for Population Pharmacokinetic Models. *Journal of the American Statistical Association*, **91**, 917–927.

[254] Wasserman, L. (1997). Bayesian Model Selection and Model Averaging. *Technical Report*, Department of Statistics, Carnegie Mellon University, USA.

[255] Wei, C.Z. (1992). On Predictive Least Squares Principles. *Annals of Statistics*, **20**, 1–42.

[256] Weiss, R.E., Wang, Y. and Ibrahim, J.G. (1997). Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors. *Biometrics*, **53**, 592–602.

[257] Yang, R. and Berger, J.O. (1996). A Catalog of Noninformative Priors. *Discussion Paper*, Institute of Statistics and Desicion Sciences, Duke University, USA.

[258] York, J., Madigan, D., Heuch, I. and Lie, R.T. (1995). Birth Defects Registered by Double Sampling: a Bayesian Approach Incorporating Covariates and Model Uncertainty. *Applied Statistics*, **44**, 227–242.

[259] Young, A.S. (1982). The Bivar Criterion for Selecting Regressors. *Technometrics*, **24**, 181–189.

[260] Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis Using G-Prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Amsterdam: North-Holland, 233–243.

[261] Zhang, P. (1997). Comment on 'An Asymptotically Theory for Linear Model Selection'. *Statistica Sinica*, **7**, 254–258.

[262] Zheng, X. and Loh, W.Y. (1995). Consistent Variable Selection for Linear Models. *Journal of the American Statistical Association*, **90**, 151–156.

[263] Zheng, X. and Loh, W.Y. (1997). A Consistent Variable Selection Criterion for Linear Models with High-Dimensional Covariates. *Statistica Sinica*, **7**, 311–325.