Bayesian Model and Variable Selection Using MCMC: Applications in Medicine

Ioannis Ntzoufras

Department of Statistics and Actuarial Science,

School of Sciences, University of the Aegean

2001, April, Skiathos

For postcript file: http://stat-athens.aueb.gr/~jbn/ntzoufras.html

Contents

Bayesian Theory

For illustration, consider the simple normal linear regression model

$$Y_i \sim N(\beta_1 X_{1i} + \beta_2 X_{2i}, \sigma^2).$$

A Bayesian model formulation consists from

1. The model likelihood $f(\underline{y}|\beta_1, \beta_2, \sigma^2)$
2. The prior distribution $f(\beta_1, \beta_2, \sigma^2)$

Then Bayesian methods are interested in estimating the posterior distributions:

- Joint posterior distribution

  $$f(\beta_1, \beta_2, \sigma^2|\underline{y}) \propto f(\underline{y}|\beta_1, \beta_2, \sigma^2) f(\beta_1, \beta_2, \sigma^2)$$

  and/or

- Marginal posterior distributions $f(\beta_1|\underline{y})$, $f(\beta_2|\underline{y})$ and $f(\sigma^2|\underline{y})$.

Markov Chain Monte Carlo Methodology

A...Gibbs Sampler

Gibbs sampler generates samples iteratively from each the full conditional posterior distributions.

- Generate β_1 from $f(\beta_1|\beta_2, \sigma, \underline{y})$
- Generate β_2 from $f(\beta_2|\beta_1, \sigma, \underline{y})$
- Generate σ^2 from $f(\sigma^2|\beta_1, \beta_2, \underline{y})$.

B...Metropolis Hastings Algorithm

In Metropolis-Hasting algorithm we

- Propose new candidate values $(\beta_1', \beta_2', \sigma^{2'})$ generated from a known easy-to-use distribution

$$q(\beta_1, \beta_2, \sigma^2, \beta_1', \beta_2', \sigma^{2'}).$$

- Accept the above proposed values with probability

$$\alpha = min\left(1, \frac{f(\beta_1', \beta_2', \sigma^{2'}|\underline{y})q(\beta_1', \beta_2', \sigma^{2'}, \beta_1, \beta_2, \sigma^2)}{f(\beta_1, \beta_2, \sigma^2|\underline{y})q(\beta_1, \beta_2, \sigma^2, \beta_1', \beta_2', \sigma^{2'})}\right)$$

Bayesian Model Selection

Bayesian model selection is based on

1. Posterior odds of model m_0 versus model m_1 given by
$$PO_{01} = \frac{f(m_0|\underline{y})}{f(m_1|\underline{y})} = \underbrace{\frac{f(\underline{y}|m_0)}{f(\underline{y}|m_1)}}_{Bayes\ Factor} \times \underbrace{\frac{f(m_0)}{f(m_1)}}_{Prior\ Odds} .$$

2. Posterior probabilities given by
$$f(m|\underline{y}) = \frac{f(\underline{y}|m)f(m)}{\sum_{m_l \in \mathcal{M}} f(\underline{y}|m_l)f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l,m}\right)^{-1},$$
   - $\mathcal{M}$: set of models under consideration,
   - $\sum_{m' \in \mathcal{M}} f(m|\underline{y}) = 1$.

Model Selection via Markov Chain Monte Carlo Methods

Problems in Bayesian model selection:

- Integrals involved in $f(m|\underline{y})$ and
- Size of $\mathcal{M}$.

Hence, MCMC methods become an extremely attractive alternative.

Descriprion:

- Generate sample $(m^{(t')}, \underline{\beta}^{(t')},\ t' = 1, \ldots, t)$
- Estimate posterior model probabilities by

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

  $I(\cdot)$: Indicator function.

- Get samples from $f(\underline{\beta}_{(m)}|m, \underline{y})$ (automatically available).

Reversible Jump
(Green, 1995, Biometrika)

<u>The Procedure</u>

Current state of the Markov chain: $\left(\underline{\beta}_{(m)}, m\right)$, then

- Generate $\underline{\beta}_{(m)}$ from $f(\underline{\beta}_{(m)}|\underline{y}, m)$ (optional).
- Propose a new model m' with probability $j(m, m')$.
- Generate $\underline{u}$ from proposal $q(\underline{u}|\underline{\beta}_{(m)}, m, m')$.
- Set $(\underline{\beta}'_{(m')}, \underline{u}') = h_{m,m'}(\underline{\beta}_{(m)}, \underline{u})$.
  - $d(m) + d(\underline{u}) = d(m') + d(\underline{u}')$ and
  - $h_{m',m} = h^{-1}_{m,m'}$.
- Accept the proposed move to model m' with probability $\alpha = \min(1, A)$

$$A = \frac{f(\underline{y}|\underline{\beta}'_{(m')}, m') f(\underline{\beta}'_{(m')}|m') f(m') j(m', m) q(\underline{u}'|\underline{\beta}'_{(m')}, m', m)}{f(\underline{y}|\underline{\beta}_{(m)}, m) f(\underline{\beta}_{(m)}|m) f(m) j(m, m') q(\underline{u}|\underline{\beta}_{(m)}, m, m')} \left| \frac{\partial h(\underline{\beta}_{(m)}, \underline{u})}{\partial(\underline{\beta}_{(m)}, \underline{u})} \right|.$$

Developments of MCMC Model and Variable Selection Methods

- New Variable Selection Samplers
  - Gibbs variable selection.
  - Metropolised Carlin and Chib sampler.
- Connections between MCMC methods
- Discussion of selection of Proposal Distributions.
- Implementation in GLM
- Link and Variable Selection
- Prior Specification

Gibbs Variable Selection

- Natural hybrid of
  - SSVS (George & McCullogh, 1993, *JASA*) &
  - Kuo and Mallick (1998, *Sankya B*) sampler.
- The model indicator m is substituted by $\underline{\gamma}^T = (\gamma_1, \ldots, \gamma_p)$ (as in SSVS).
- The vector of $\underline{\gamma}$ participates in likelihood via the linear predictor:

$$\eta_i = \sum_{j=1}^{p} \underline{X}_{ij} \gamma_j \underline{\beta}_j .$$

- Specify prior as $f(\underline{\beta}|\underline{\gamma}) f(\underline{\gamma})$.
  Consider the partition of $\underline{\beta} = \left\{ \underline{\beta}_{(\underline{\gamma})}, \underline{\beta}_{(\backslash\underline{\gamma})} \right\}$ into
  - $\underline{\beta}_{(\underline{\gamma})}$: parameters in model ($\gamma_j = 1$)
  - $\underline{\beta}_{(\backslash\underline{\gamma})}$: parameters not in model ($\gamma_j = 0$)

  then $f(\underline{\beta}|\underline{\gamma})$ may be partitioned into
  - Prior: $f(\underline{\beta}_{(\underline{\gamma})}|\underline{\gamma})$ and
  - Pseudoprior: $f(\underline{\beta}_{(\backslash\underline{\gamma})}|\underline{\beta}_{(\underline{\gamma})}, \underline{\gamma})$.

The Procedure

Suppose that the current state is $(\underline{\beta}, \underline{\gamma})$, then

- Generate parameters $\underline{\beta}_{(\underline{\gamma})}$ from

$$f(\underline{\beta}_{(\underline{\gamma})}|\underline{\beta}_{(\backslash\underline{\gamma})}, \underline{\gamma}, \underline{y}) \propto f(\underline{y}|\underline{\beta}, \underline{\gamma}) f(\underline{\beta}_{(\underline{\gamma})}|\underline{\gamma}) f(\underline{\beta}_{(\backslash\underline{\gamma})}|\underline{\beta}_{(\underline{\gamma})}, \underline{\gamma})$$

- Generate pseudo-parameters $\underline{\beta}_{(\backslash\underline{\gamma})}$ from

$$f(\underline{\beta}_{(\backslash\underline{\gamma})}|\underline{\beta}_{(\underline{\gamma})}, \underline{\gamma})$$

- Generate $\gamma_j \sim Bernoulli\left(\frac{O_j}{1+O_j}\right)$ with

$$O_j = \underbrace{\frac{f(\underline{y}|\underline{\beta}, \gamma_j = 1, \underline{\gamma}_{\backslash j})}{f(\underline{y}|\underline{\beta}, \gamma_j = 0, \underline{\gamma}_{\backslash j})}}_{Likelihood\ Ratio} \underbrace{\frac{f(\underline{\beta}|\gamma_j = 1, \underline{\gamma}_{\backslash j})}{f(\underline{\beta}|\gamma_j = 0, \underline{\gamma}_{\backslash j})}}_{Prior/Pseudoprior\ Ratio} \underbrace{\frac{f(\gamma_j = 1, \underline{\gamma}_{\backslash j})}{f(\gamma_j = 0, \underline{\gamma}_{\backslash j})}}_{Prior\ Odds}$$

Coronary Heart Disease Risk Factors: Edwards and Havránek (1985, Biometrika)

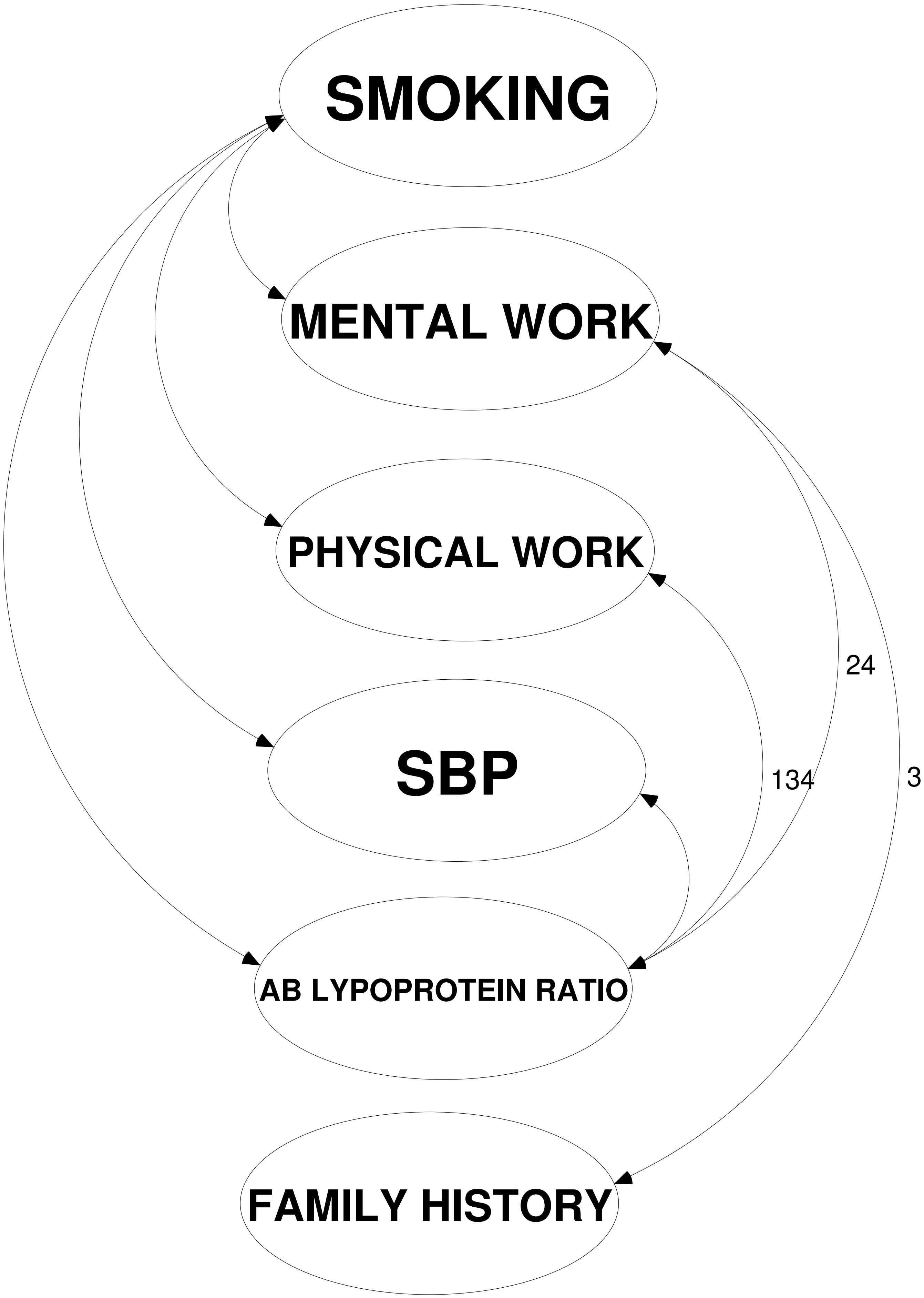
1841 men cross-classified according to six factors of Coronary heart disease :

- 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.

Approximately 7.8 million hierarchical log-linear models!!!

Model	Posterior Probability	
	RJ	GVS
AC+BC+AD+AE+CE+DE+F	0.260	0.270
AC+BC+AD+AE+BE+DE+F	0.166	0.154
AC+BC+AD+AE+CE+DE+BF	0.075	0.077
AC+BC+AD+AE+BE+CE+DE+F	0.063	0.069
Other Models	0.460	0.430
Mean of Batch Standard Deviation	0.064	0.025

Table 1: 2^6 Contingency Table: Posterior Model Probabilities.


SMOKING
MENTAL WORK
PHYSICAL WORK
SBP
AB LYPOPROTEIN RATIO
FAMILY HISTORY
24
134
3

The Effect of Diet on Thyroid Cancer

Details

- Work with Ass. Prof. A.Linos, Dr. D.Linos and I.Markaki.
- Data from Patients in Various Hospitals in Athens in Early 90's.
- Assess the relationship between diet and thyroid disease, either malignant or benign.
- Diet was recorded with the use of a Food Frequency Questionnaire (servings per month).

Application

Perform Logistic Regression Analysis with:

- Response:
  - Thyroid Cancer Cases vs. Controls
  - Benign Thyroid Disease Cases vs. Controls
- Explanatory Variables: Age, Gender, Body Mass Index -BMI(kgr/m^2)- and "dietary habits" expressed by
  - 31 food items (2,147,483,648 Models) or
  - 13 food patterns (8,192 Models)

Comments

- Dietary patterns: PCA (VARIMAX rotation)
- We consider components with eigenvalues> 1
- 62.3% of the total variance explained

	Thyroid Ca vs. Controls	Benign Thyroid vs. Controls
Food Item	**Posterior Probabilities**	**Posterior Probabilities**
Leafy Vegetables*	80%	-
Tomatoes	65%	36%
Broccoli*	64%	87%
Lentils	62%	64%
Pork	61%	-
Eggs	60%	-
***Salted snacks** +*	58%	31%
Margarine	54%	-
Spinach*	52%	66%
Lemons	52%	-
Carrots*	51%	-
Onions*	-	92%

Conclusions For Thyroid Cancer

- Analysis of Food Items
  - Most Important Protective Items: Leafy Vegetables, Tomatoes and Broccoli.
  - Most Important Causative Items: Lentils, Pork and Eggs.
  - Lowest Risk: Broccoli (OR=0.74, PP=0.64).
  - Highest Risk: Lentils (OR=1.46, PP=0.64) and Spinach (OR=1.25, PP=0.52)
- Analysis of Food Patterns
  - Lowest Risk: 2nd Component (tomatoes, onions, lemons and cauliflower, OR=0.55, PP=0.94).
  - Highest Risk: 9th Component (Butter and Eggs, OR=1.53, PP=0.64)

Discussion

- The Problem of Model Selection is Essentially Unsolved and Subjective.

- MCMC can Help Estimating Posterior Probabilities and Applying Bayesian Model Averaging.

- Bayesian Model Selection for GLM Can be Easily Applied in an Automatic Pattern.

- Bayesian Model Averaging is even More Important for Medical Data (see for example for survival analysis: Raftery *et al.* , 1996 VAL5, Volinsky *et al.* , 1997, JRSSC).

Special Thanks:

- My Supervisors:
  - Petros Dellaportas and
  - Jonathan J. Forster.
- My parents