MIXTURES EVERYWHERE

Dimitris Karlis and Evdokia Xekalaki
Department of Statistics
Athens University of Economics & Business, Greece

1. Introduction to Mixture Models

Mixture models are widely used in statistical modeling since they can model situations which a simple model cannot adequately describe. In recent years, mixture modeling has been exploited mainly due to high-speed computers that can make tractable problems that occur when working with mixtures (e.g. estimation). Statistics has benefited immensely by the development of advanced computer machines and thus more sophisticated and complicated methodologies have been developed. Mixture models underlie the use of such methodologies in a wide spectrum of practical situations where the hypothesized models can be given a mixture interpretation as demonstrated in the sequel.

In general, mixtures provide generalizations of simple models. For example, assuming a specific form for the distribution of the population that generated a data set implies that the mean to variance relation is given for this distribution. In practical situations this may not always be true. A simple example is the Poisson distribution. It is well known (see, e.g., Johnson et al*.,* 1992) that for the Poisson distribution the variance is equal to the mean. Hence, assuming a Poisson distribution implies a mean to variance ratio equal to unity. With real data sets however, this is rarely the case. Quite often, the sample mean is noticeably exceeded by the sample variance. This situation is known as *overdispersion*. A Poisson distribution is no longer a suitable model in such a case and the need of a more general family of distributions becomes obvious. Such a flexible family may be defined if one allows the parameter (or the parameters) θ of the original distribution to vary according to a distribution with probability density function, say $g(\cdot)$.

Definition 1. A distribution function $\mathrm{F}(\cdot)$ is called *a mixture of the distribution function* $\mathrm{F}(\cdot \mid \theta)$ *with mixing distribution* $\mathrm{G}_\theta(\cdot)$ if it can be written in the form

$$\mathrm{F_x(x)} = \int_\Theta \mathrm{F_{x|\theta}(x \mid \theta) dG_\theta(\theta)}\,, \qquad (1)$$

where Θ is the space in which θ takes values and $G_\theta(\cdot)$ can be continuous, discrete or a finite step distribution.

The above definition can be also expressed in terms of probability density functions in the continuous case (or the probability functions in the discrete case). The above mixture is denoted as $\mathrm{F_{x|\theta}(x)} \underset{\theta}{\wedge} \mathrm{G}(\theta)$. In the sequel, a mixture with a finite step mixing distribution will be termed *a k-finite step mixture* of $F(\cdot \mid \theta)$, where k is a non-negative integer referring to the number of points with positive probabilities in the mixing distribution.

Mixture models cover several distinct fields of the statistical science. Their broad acceptance as plausible models in diverse situations is reflected in the statistical literature. Titterington et al. (1985) provide an extensive review of work in the area of mixture models up to 1985. In recent years, the number of applications increased mainly because of the availability of high speed computer resources. Moreover, since many methods can be seen through the prism of mixture models, there is a vast literature concerning applications of mixture models in various contexts. Recent reviews on mixtures can be found in Lindsay (1995), Bohning (1999), McLachlan and Peel (2001).

The purpose of this paper is to bring together various models from diverse fields that are in fact mixture models. The resulting collection of models may be far from being exhaustive as the focus has been on methodologies that are common in statistical practice and not on results concerning special cases. To this extent, the number of articles cited was kept to a minimum and reference was made only to a selection of papers that could pilot the reader in the various areas.

In Section 2 of the chapter two basic concepts are discussed in the context of which the mixture models are used: overdispersion and inhomogeneity. Section 3 presents various statistical methodologies that use the idea of mixtures. An attempt is made to show clearly the connection of such methodologies to mixture models. Finally, in Section 4 a brief discussion is provided highlighting the implications of a unified treatment of all the models discussed.

2. General Properties

2.1 Inhomogeneity models

Mixture models are used to describe inhomogeneous populations. The *i*-th group of individuals of the population has a distribution defined by a probability density function $f(\cdot|\theta_i)$. All the members of the population follow the same parametric form of distribution, but the parameter θ_i varies from individual to individual according to a distribution $G_\theta(\cdot)$. For example, considering the number of accidents incurred by a population of clients of an insurance company, it is reasonable to assume that there are at least two subpopulations, the new drivers and the old drivers. Drivers can thus be assumed to incur accidents at rates that differ from one subpopulation to the other subpopulation, say $\theta_1 \neq \theta_2$.This is the simplest form of inhomogeneity: the population consists of two subpopulations. Allowing for the number of subpopulations to tend to infinity, i.e., considering different categories of drivers according to infinitely many characteristics, such as age, sex, origin, social, and economic status, etc. a continuous mixing distribution for the parameter θ of the Poisson distribution arises.

Depending on the choice of the mixing distribution $G_\theta(\cdot)$, a very broad family of distributions is obtained, which may be adequate for cases where the simple model fails. So, a mixture model describes an inhomogeneous population while the mixing distribution describes the inhomogeneity of the population. If the population were homogeneous, then all the members would have the same parameter θ, and the simple model would adequately describe the situation.

2.2 Overdispersion

A fundamental property of mixture models stems from the following representation of the variance of the mixed variate X.

$$Var(X) = Var(E(X\,|\,\theta)) + E(Var(X\,|\,\theta))\,. \quad (2)$$

The above formula separates the variance of X into two parts. Since the parameter θ represents the inhomogeneity of the population, the first part of the variance represents the variance due to the variability of the parameter θ, while the second part reflects the inherent variability of the random variable X if θ did not vary. One can recognize that a similar idea is the basis for ANOVA models where the total variability is split into the "between groups" and the "within groups" components. This is further discussed in Section 3.

The above formula offers an explanation as to why mixture models are often termed as *overdispersion models*. A mixture model has a variance greater than that of the simple model (e.g., Shaked, 1980). Thus, it is commonly proposed that if the simple model cannot describe the variability present in the data, overdispersed alternatives based on mixtures could be used.

3. Fields of Application

3.1 Data Modelling

The main advantage of mixture models lies in that they provide the possibility of generalizing existing simple models through an appropriate choice of a mixing distribution which acts as a means of "loosening" the structure of the initial model by allowing its parameter to vary. A wealth of alternative models can thus be considered whenever the simple (initial) model fails and many interesting distributions may be obtained from simple and well-known distributions such as the Poisson, the binomial, the normal, the exponential, through mixing procedures.

In recent years, the computational difficulties for applying such complicated models have disappeared and some new distributions (discrete or continuous) have been proposed. Moreover, since mixture models are widely used to describe inhomogeneous populations they have become a very popular choice in practice, since they offer realistic interpretations of the mechanisms that generated the data.

The derivation of the negative binomial distribution, as a mixture of the Poisson distribution with a gamma distribution as the mixing distribution, originally obtained by Greenwood and Yule (1920) constitutes a typical example. Almost all the well-known distributions have been generalized by considering mixtures of them. A large number of Poisson mixtures have been developed. (For an extensive review, see Karlis, 1998).

Perhaps, the beta binomial distribution (see, e.g., Tripathi et al., 1994) is the most famous example of binomial mixtures. Alternative models have been described in Alanko and Duffy (1996) and Brooks et al. (1997).

Negative binomial mixtures have also been widely used with applications in a large number of fields. These include the Yule distribution (Yule, 1925, Simon, 1955, Kendall, 1961, Xekalaki, 1983a, 1984b) and the generalized Waring distribution (Irwin, 1963, 1968, 1975, Dacey, 1972, Xekalaki, 1983b, 1984a). Note that negative binomial mixtures can be seen as Poisson mixtures as well.

Normal mixtures on the parameter representing the mean of the distribution are not common in practice. Mixtures of the normal distribution on the parameter representing its variance are referred to as *scale mixtures* (e.g., Andrews and Mallows, 1974). For example, the t-distribution is a scale mixture of the normal distribution with a chi-square mixing distribution. Barndorff-Nielsen et al. (1982) described a more general family of normal mixtures of the form $f_{N(\mu+\theta\beta,\theta a)} \underset{\theta}{\wedge} g(\theta)$, where $f_{N(\alpha,\beta)}$ stands for the probability density function of the normal distribution with mean α and variance β. The distributions arising from such mixtures are not necessarily symmetric and have heavier tails than the normal distribution. Applications of normal scale mixtures have been considered by Barndorff-Nielsen (1997) and Eberlein and Keller (1995).

Similarly, exponential mixtures are described in Hebert (1994) and Jewell (1982), for life testing applications. The beta distribution can be seen as a Gamma mixture, while the Gamma distribution can be seen as a scale mixture of the exponential distribution (Gleser, 1989). Many other mixture distributions have been proposed in the literature.

A wide family of distributions can be defined to consist of finite mixtures of distributions, with components not necessarily from the same family of distributions. Finite mixtures with different component distributions have been described in Rachev and Sengupta (1993) (Laplace - Weibull), Jorgensen et al. (1991) (Inverse Gaussian – Reciprocal Inverse Gaussian), Scallan (1992) (Normal – Laplace), Al-Hussaini and Abd-El-Hakim (1989) (Inverse Gaussian-Weibull) and many others.

Finally, note that mixture models can have a variety of shapes that are never taken by simple models, such as multimodal shapes. These are usually represented via finite mixtures. So, for example, mixing two normal distributions of equal variances in equal proportions can result in a bimodal distribution with well-separated modes, appropriate for describing data exhibiting such a behavior.

3.2 Discriminant Analysis

In discriminant analysis, one needs to construct rules so as to be able to distinguish the subpopulation from which a new observation comes. Assuming a finite mixture model one may obtain the parameters of the subpopulations from a training set, and then classify the new observations via simple probabilistic arguments (see, e.g., McLachlan, 1992). This approach is also referred to as *statistical pattern recognition* in computer science applications.

Consider a population consisting of k subpopulations, each distributed according to a distribution defined by a density function $f_j(\cdot\,|\,\theta_j), j=1,2,\ldots, k$. Suppose further that the size of each subpopulation is p_j . Usually, data used in discriminant analysis also contain variables Z_j , $j=1,2, \ldots, k,$ which take the value 1 if the observation belongs to the j-th subpopulation and 0 otherwise. These data are used for estimating the parameters θ_j , p_j and are referred to as *training data*. Then, a new observation x is allocated to each group according to its posterior probability of belonging to the j-th group

$$P(Z_j = 1\,|\,x) = \frac{p_j f_j(x\,|\,\theta_j)}{\sum_{j=1}^{k} p_j f_j(x\,|\,\theta_j)} .$$

One can recognize the mixture formulation in the above formula, as well as the fact that this formulation comprises the E-step of the EM algorithm for estimation in finite mixture models. The variables Z_{ij} are the "*missing*" data in the construction of the EM algorithm for finite mixtures.

However, such data sets often contain a lot of unclassified observations, i.e., observations that do not relate to specific values of Z_j , $j=1,2, \ldots, k$ and hence one can use these data for estimation purposes. The likelihood function for such data is expressed in terms of the mixture $\sum_{j=1}^{k} p_j f_j(x\,|\,\theta_j)$ and standard mixture methodologies must be used for estimating the parameters. Note that unclassified observations contribute to the estimation of all the parameters (see, e.g., Hosmer, 1973).

Usually, the densities $f_j(\cdot\,|\,\theta_j)$ are assumed multivariate normal with both the mean vector and the variance-covariance matrix being variable.

It is interesting to note that although the EM algorithm for mixtures was introduced quite early by Hasselblad (1969), it did not find wide usage until computer machines became widely available. This typically reflects the impact of computer resources in mixture modeling. The same is true of a wide range of fields that, despite their early development, attracted greater interest only after the generalized use of statistical software. Cluster analysis is another typical example.

3.3 Cluster Analysis

Finite mixtures play an important role to the development of methods in cluster analysis. Two main approaches are used for clustering purposes. The first considers distances between the observations and then clusters the data according to their distances from specific cluster centers. The second approach utilizes a finite mixture model.

The idea is to describe the entire population as a mixture model consisting of several subpopulations (clusters). Then, a methodology could be to fit this finite mixture model and subsequently use the estimated parameters to obtain the posterior probability with which each of the observations belongs to the j-th subpopulation (McLachlan & Basford, 1989). According to a decision criterion, each observation is allocated to a subpopulation, thus creating clusters of data. The problem of choosing the number of clusters that best describe the data, reduces to that of selecting the number of support points for the finite mixture (see, e.g., Karlis & Xekalaki, 1999).

Usually, multivariate normal subpopulations are considered (Banfield & Raftery, 1993 and McLachlan & Basford, 1989). Symons et al. (1983) found clusters of Poisson distributed data for an epidemiological application, while data containing both continuous and discrete variables can be analyzed via multivariate normal densities where thresholds are used for the categorical variables (see, e.g., Everitt & Merette, 1990).

3.4 Outlier-robustness Studies

Outliers in data sets have been modelled by means of mixture models (see, e.g., Aitkin & Wilson, 1980). It is assumed that an outlier comprises a component in a mixture model. More formally, the representation used for the underlying model is

$$(1-p)\ f(\cdot\,|\,\theta)\ + p\,g(\cdot),$$

where $f(\cdot\,|\,\theta)$ is the true density *contaminated* by a proportion of p observations from a density $g(\cdot)$. Hence, by fitting a mixture model we may investigate the existence of outliers. In robustness studies, the contamination of the data can also be regarded as an additional component of a mixture model.

In addition, for robustness studies with normal populations it is natural to use a t-distribution. Recall that the t-distribution is in fact a scale mixture of the normal distribution. Other scale mixtures have also been proposed for examining robustness of methods for normal populations (e.g., Cao & West, 1996).

Note further, that since mixtures of a distribution tend to this distribution if a degenerate mixing distribution is used, it would be natural to consider the general mixture model as leading to the simple model as the variance of the underlying model decreases.

3.5 Analysis of Variance (ANOVA) Models

The wellknown technique of the analysis of variance is a particular application of mixture models. It is assumed that the mean of the normal distribution of the entire population, varies from subpopulation to subpopulation and the total variance is decomposed with respect to randomness and mixing.

The simple ANOVA model assumes prespecified values for the means of the different components, not allowing them to vary. The case where the means come from a distribution with density $g(\cdot)$ corresponds to the so-called *random effects model* described in the sequel. It is interesting that the simple ANOVA models are based on the inhomogeneity model which allows for subpopulations with different means. The decomposition of the variance given in (2) is the classical ANOVA model separating the total variance into the "between groups" variance and the "within groups" variance.

Beyond the widely applied classical ANOVA model for normal populations, similar ANOVA models have been proposed for discrete data as well. For example, Irwin (1968) and Xekalaki (1983b, 1984a), in the context of accident theory, considered analyzing the total variance into three additive components corresponding to internal and external non-random factors and to random factors. Also, Brooks (1984) described an ANOVA model for beta-binomial data.

3.6 Random Effects Models and Related Models

Consider the classical one-way ANOVA model. It is assumed that the i-th observation of the j-th group, say X_{ij} , follows a $N(\theta_j,\sigma^2)$ distribution, where θ_j is the mean of the j-th group. The simple ANOVA model assumes that the values of θ_j*'s* are prespecified. Random effect models assume that the parameters are not constant but they are realizations from a distribution with density $g(\cdot)$. The resulting marginal density function of the data for each group is of the form

$$f(\mathbf{x}_j) = \int_{\mathbf{\Theta}} \prod_{i=1}^{n_j} f(x_{ij};\theta,\sigma^2) g(\theta) d\theta ,$$

where n_j is the sample size of the j-th subpopulation. The usual choice for $g(\cdot)$ is the density function of the normal distribution, resulting in normal marginals. This choice was based mainly on its computational tractability, since other choices led to complicated marginal distributions.

Such random effects models have been described for the broad family of Generalized Linear Models. Consider, for example, the Poisson regression case. For simplicity, we consider only a single covariate, say X. A model of this type assumes that the data Y_i , $i=1, 2, \ldots ,n$ follow a Poisson distribution with mean λ_i such that

$$\log(\lambda_i) = a + \beta X_i + \varepsilon_i$$

for some constants a,β and with ε_i having a distribution with mean equal to 0 and variance say φ. Now the marginal distribution of the observations y_i is no longer the Poisson distribution, but a mixed Poisson distribution, with mixing distribution clearly depending on the distribution of ε_i.

From the regression equation, one can obtain that

$$Y_i \sim Poisson(t_i \exp(a + \beta X_i)) \underset{t}{\wedge} g(t) ,$$

where $t_i = \exp(\varepsilon_i)$ with a distribution that depends on the distribution of ε_i. Negative Binomial and Poisson Inverse Gaussian regression models have been proposed as overdispersed alternatives to the Poisson regression model (Lawless, 1987, Dean et al., 1989). If the distribution of t is a two finite step distribution, the finite Poison mixture regression model of Wang et al. (1996) results. The similarity of the mixture representation and the random effects one is discussed in Hinde and Demetrio (1998).

The above example from the Poisson distribution can easily be generalized. All the random effect models introduce a mixing distribution for the error which adds one more term to the total variability.

Very similar to the random effect model is the formulation of repeated measurement models, where the added variability is due to the variance induced by the series of observations on the same individual. Starting from a linear regression (in general one can consider any link function suitably linearized) one can add variability by regarding any term of the linear regression model as a random variable. So, allowing the intercept parameter to vary leads to *random coefficient regression models* (see Mallet, 1986). Also, *error-in-variables models* arise by letting the covariates themselves vary. Finally, *random effect models* are obtained if the error term is allowed to vary.

Lee and Nelder (1996) discussed the above models under the general caption of *hierarchical generalized linear models*.

3.7 Kernel Density Estimation

In *kernel density estimation*, the aim is to estimate a probability density function $f(.)$ on the basis of a sample of size *n* by smoothing the probability mass of $\frac{1}{n}$ placed at each of the observations (x_i) by the empirical distribution function according to a kernel $K_n(\cdot, x_i)$. This is usually a symmetric probability density function of the form $K_n(x, x_i) = K\left(\frac{x - x_i}{h}\right)$, $i = 1,\ 2,\ ...,\ n$, where h is a switching parameter which handles the smoothing procedure (see, e.g., Silverman, 1986). Thus, in kernel density estimation a kernel mixture model is considered with equal mixing probabilities. More specifically, the density estimate $\hat{f}_n(.)$ of $f(.)$ at the point x is obtained by

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

One can recognize that the above representation of the kernel estimate is a n-finite mixture of the kernel $K(\cdot, \cdot)$. Though this mixture representation has been recognized (see, e.g., Simonoff, 1996), it has not been exploited in practice. The idea is to use certain kernels in order to obtain estimates with useful properties. For example, data restricted on the positive axis can be estimated using exponential or gamma kernels (see, e.g., Chen, 2000), depending on the shape (J-shaped or bell shaped data). Similarly, discrete data can be estimated via Poisson kernels etc.

Moreover, specific approaches can be used in order to achieve certain smoothing properties. By using such approaches the choice of the smoothing parameter *h* can be reduced to the choice of the kernel parameters so that the smoothing properties be fulfilled. Wang and Van Ryzin (1979) described such an approach in an empirical Bayesian context for the estimation of a Poisson mean. They proposed estimating the discrete density, given the data $X_1, X_2, ..., X_n$ by

$$\hat{P}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{\exp(-x_i)x_i^x}{x!}$$

i.e., as a mixture of *n* Poisson distributions with parameters equal to the observations x_i, i=1, 2, …, n.

3.8 Latent Structure Models and Factor Analysis

In latent structure models it is assumed that beyond the observable random variables there are other unobservable or even non-measurable variables, which influence the situation under investigation. The main assumption in the case of latent structure models is that of conditional independence, i.e., the assumption that for a given value of the unobservable variable the remaining variables are independent. Since inference is based on the unconditional distribution, we obtain, by the law of total probability, a mixture model where the mixing distribution represents the distribution of the unobservable quantity which thus is of special interest in many situations (see, e.g., Everitt, 1984). It is very interesting that many methods proposed for mixture models are applicable to latent variable models (see, e.g., Aitkin et al., 1981).

For example, in psychological tests the probability that the i-th person will correctly answer x questions is described as $p(x|\varphi_i)$ where φ_i represents the ability of the i-th person. Additionally, it is assumed that given the ability of each person the scores x are independent. (This is the idea of conditional independence). Since, ability is a rather abstract notion that cannot be measured, the researcher may assume either a parametric form of distribution for its values (e.g., a normal distribution with some parameters) or a finite step distribution (as in the case where φ_i can take only a finite number of different values). This has a common element with the method of factor analysis for continuous variables (Bartholomew, 1980).

A formulation of the problem is the following. Suppose that one observes a set of *p*-variables, say $\mathbf{x} =(x_1,x_2, \ldots., x_p)$. A latent structure model supposes that these variables are related to a set of *q* unobservable and perhaps non-measurable variables (e.g., some abstract concepts such as hazard, interest, ability, love, etc.), say $\mathbf{y} = (y_1,y_2, \ldots., y_q.)$. For the model to be practically useful, *q* needs to be much smaller than *p*. The relationship between **x** and **y** is stochastic and may be expressed by a conditional probability function $\pi(\mathbf{x}|\mathbf{y})$ being the conditional distribution of the variables **x** given the unobservable **y**. The purpose of latent structure models is to infer on **y**, keeping in mind, that we have observed only **x**. The marginal density of **x** can be represented as a mixture, by

$$f(\mathbf{x}) = \int \pi(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} .$$

One can infer on **y** using the Bayes theorem, since

$$\pi(\mathbf{y} \mid \mathbf{x}) = \frac{\pi(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y})}{f(\mathbf{x})} .$$

Hence, the problem reduces to one of estimating the mixing density $\mathrm{p}(\cdot)$. As described earlier, this density can be either specified parametrically, and hence only the parameters of the defined density must be estimated, or it can be estimated non-parametrically (see, e.g., Lindsay et al., 1991).

Latent structure models can be considered as factor analysis models for categorical data. The classical factor analysis model assumes that a set of observable variables, say **x** $=(x_1,x_2, \ldots., x_p)$ can be expressed as a linear combination of a set of unobservable variables, say **y** = $(y_1,y_2, \ldots., y_q.)$, termed *factors*. More formally

$$\mathbf{x} = \mathbf{B}\mathbf{y} + \boldsymbol{\varepsilon}$$

where the matrix **B** contains the *factor loadings*, i.e., its (i, j) element is the contribution of the j-th factor to the determination of the i-th variable. The vector of errors **ε** contains the unexplained part of each variable and it is assumed to follow a *N(0,***D***)*, where $\mathbf{D} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$. Conditionally on the factors **y**, **x|y** ~ **N(By,D)** distribution, and the factors follow themselves a **N(0,I**$_q$) distribution. Then, the unconditional distribution of **x** is a **N(0, BB**^{t}**+D)** distribution. Note that the variance-covariance matrix is decomposed into two terms, the variance explained by the factors and the remaining unexplained part. This decomposition is the basis for the factor analysis model.

3.9 Bayes and Empirical Bayes Estimation.

Bayesian statistical methods have their origin in the well-known Bayes theorem. From a Bayesian perspective, the parameter θ of a density function, say $f(\cdot \mid \theta)$ has itself a distribution function $\mathrm{g}(\cdot)$, termed the *prior*, reflecting one's belief about the parameter and allowing for extra variability. We treat θ as a scalar for simplicity, but clearly, it can be vector valued as well. The prior distribution corresponds to the mixing distribution in (1).

The determination of the prior distribution is crucial for the applicability of the method. Standard Bayesian methods propose a prior based on past experience, on the researcher's belief, or a non-informative prior in the sense that no clear information about the parameter exists and this ignorance is accounted for by a very dispersed prior. Instead of determining the prior by specific values of its parameters, recent hierarchical Bayes models propose treating the parameters of the prior distribution as random variates and imposing hyperpriors on them. Such an approach can remove subjectivity with respect to the selection of the prior distribution.

A different approach is that of the so-called Empirical Bayes methodologies (see, e.g., Karlin & Lewis, 1996). Specifically, the Empirical Bayesian methods aim at estimating the prior distribution from the data. This reduces to the problem of estimating the mixing distribution. This obvious relationship between these two distinct areas of statistics have resulted in a vast number of papers in both areas, with many common elements (see, for example, Maritz & Lwin, 1989, Laird, 1982). The aim is the same in both cases, though the interest lies in different aspects.

Putting aside the relationship of Bayesian approaches and mixture models, there are several other topics in the Bayesian literature that use mixtures in order to improve the inferences made. For example, mixtures have been proposed to be used as priors, the main reason being their flexibility (see, e.g., Dalal & Hall, 1983). Beyond that, such priors are also robust and have been proposed for examining Bayesian robustness (Bose, 1994). Escobar and West (1995) proposed mixtures of normals as an effective basis for nonparametric Bayesian density estimation.

3.10 Random Variate Generation

The mixture representation of some distributions is a powerful tool for efficient random number generation from these distributions. Several distributions (discrete or continuous) may arise as mixture models from certain distributions, which are easier to generate. Hence, generating variables in the context of such a representation can be less expensive.

For example, variables can be generated from the negative binomial distribution by utilizing its derivation as a mixture of the Poisson distribution with a Gamma mixing distribution. Another, more complicated example of perhaps more practical interest is given by Philippe (1997). She considered generation of truncated gamma variables based on a finite mixture representation of the truncated Gamma distribution.

Furthermore, the distributions of products and ratios of random variables can be regarded as mixtures and hence the algorithms used to simulate from such distributions are in fact examples of utilizing their mixture representation. For more details, the reader is referred to Devroye (1992).

3.11 Approximating the Distribution of a Statistic

In many statistical methods, the derived statistics do not have a standard distributional form and an approximation has to be considered for their distribution. Mixture models allow for flexible approximation in such cases. Such an example is the approximation of the distribution of the correlation coefficient used in Mudholkar and Chaubey (1976). In order to cope with the inappropriateness of the normal approximation of the distribution of the sample correlation coefficient, especially in the tails of the distribution, they proposed the use of a mixture of a normal distribution with a logistic distribution. Such a mixture results in a distribution with heavier tails suitable for the distribution of the correlation coefficient

3.12 Multilevel Models

Multilevel statistical models assume that one can separate the total variation of the data into levels and estimate the component attributed to each level (see, e.g., Goldstein, 1995). Consider a k-level model and let (y_{ij}, x_{ij}) denote the i-th observation from the j-th level. In the context of the typical linear model

$$y_{ij} = a_j + b_j x_{ij} + e_{ij}$$

one has to estimate the parameters a_j, b_j, $j=1, \ldots, k$ and the variance σ^2 of the data.

A multilevel statistical model treats the parameters a_j, b_j as random variables in the sense that $a_j = c_0 + u_j$, and $b_j = c_1 + v_j$ where (u_j, v_j) follows a bivariate normal distribution with zero means and a variance covariance matrix. Then, the simple model can be rewritten as

$$y_{ij} = c_o + c_1 x_{ij} + (u_j + v_j x_{ij} + e_{ij}) .$$

A variance component is added corresponding to each level. For this reason, the model is also termed as the *variance components model*. Since normal distributions are usually used (mainly for convenience) the resulting distributions are also normal. The mixture representation is used for applying an EM algorithm for the estimation of the parameters (Goldstein, 1995).

3.13 Distributions Arising out of Methods of Ascertainment

When an investigator collects a sample of observations produced by nature according to some model, the original distribution may not be reproduced due to various reasons. These include partial destruction or enhancement of observations. Situations of the former type are known in the literature as *damage models* while situations of the latter type are known as *generating models*. The distortion mechanism is usually assumed to be manifested through the conditional distribution of the resulting random variable Y given the value of the original random variable X. As a result, the observed distribution is a distorted version of the original distribution obtained as a mixture of the distortion mechanism. In particular, in the case of damage,

$$P(Y=r)=\sum_{n=r}^{\infty} P(Y=r \mid X=n)\, P(X=n), \qquad r=0,1,.....,,$$

while in the case of enhancement

$$P(Y=r)=\sum_{n=1}^{r} P(Y=r \mid X=n)\ P(X=n), \quad r=1,2,...$$

Various forms of distributions have been considered for the distortion mechanisms in the above two cases. In the case of damage, the most popular forms have been the binomial distribution (Rao, 1963), mixtures on p of the binomial distribution (e.g., Panaretos, 1982, Xekalaki & Panaretos, 1983) whenever damage can be regarded as additive (Y=X–U, U independent of Y) or in terms of the uniform distribution in (0, x) (e.g., Xekalaki, 1984b) whenever damage can be regarded as multiplicative (Y= [RX], R independent of X and uniformly distributed in (0, 1)). The latter case has also been considered in the context of continuous distributions by Krishnaji (1970b). The generating model was introduced and studied by Panaretos (1983).

3.14 Other Models

It is worth mentioning that several other problems in the statistical literature can be seen through the prism of mixture models. For example, *deconvolution* *problems* (see, e.g., Caroll & Hall, 1988, Liu & Taylor, 1989) assume that the data X can be written as Y+Z, where Y is a latent variable and Z has a known density *f*. Then, the density of X can be written in a mixture form, thus

$$g(x) = \int f(x-y)dQ(y),$$

where $Q(\cdot)$ is the distribution function of the latent variable Y. The above model can be considered as a *measurement error model*. In this context, the problem reduces to estimating the mixing distribution from a mixture. Similar problems related to hidden Markov models are described in Leroux (1992).

Simple convolutions can be regarded as mixture models. Also, as already mentioned in Section 3.10, products of random variables can be regarded as mixture models (Sibuya, 1979).

Another application of mixtures is given by Rudas et al. (1994) in the context of testing for the goodness of fit of a model. In particular, they propose treating the model under investigation as a component in a 2-finite mixture model. The estimated mixing proportion together with a parametric bootstrap confidence interval for this quantity can be regarded as evidence for or against the assumed model. The idea can be generalized to a variety of goodness of fit problems, especially for non-nested models.

From this, it becomes evident that a latent mixture structure exists in a variety of statistical models, often ignored by the researcher. Interesting reviews for the mixture models are given in the books by Everitt and Hand (1981), Titterington et al. (1985), McLachlan and Basford (1989), Lindsay (1995), Bohning (1999), McLachlan and Peel (2001) as well as in the review papers of Gupta and Huang (1981), Redner and Walker (1984) and Titterington (1990).

4. Discussion

An anthology of statistical methods and models directly or indirectly related to mixture models was given. In some of them, the mixture idea is often well hidden in the property that a location mixture of the normal distribution is itself a normal distribution. So, the standard normal theory still holds and estimation is not difficult under a normal distribution.

A question that naturally arises is what one can gain by such mixture representations of all these models. As has been demonstrated, beyond the philosophical issue of a unified statistical approach, some elements common in all these models can be brought about. Many of these models have a structure that is of a latent nature such as containing, unobserved quantities that are not measurable, but nevertheless play a key-role in the model.

All the models discussed imply the two basic concepts of inhomogeneity and overdispersion. Further, any mixture model admits an interesting missing data interpretation. Thus, a unifying approach in modeling different situations allows the application of methodologies used in the case of mixtures to other models. Mixture models, for instance, provide the rationale on which the estimation step of the well-known EM-algorithm is based for the estimation of the unknown values of the parameters which are treated as "missing data." For example, Goutis (1993) followed an EM algorithmic approach for a logistic regression model with random effects. In general, such EM algorithms for random effect models can reduce the whole problem to one of fitting a generalized linear model to the simple distribution; such procedures are provided in standard statistical packages. Hence, iterative methods that provide estimates can be constructed. Other techniques can also be applied, like nonparametric estimation. (See, Lindsay, 1995, for an interesting elaboration). Such approaches reduce the need for specific assumptions when applying the models leading to more widely applicable models.

As mentioned in the introduction, the impact of computer resources on the development of mixture models and on the enhancement of their application potential has been tremendous. Although early work on mixtures relied on computers (e.g., the development of the EM algorithm for finite mixture models by Hasselblad, 1969, and the first attempt for model based clustering by Wolfe, 1970), the progress in this field was rather slow until the beginning of last decade. The implementation of Lindsay's (1983) general maximum likelihood theorem for mixtures, a milestone in mixture modeling, relies on computers. Non-parametric maximum likelihood estimation of a mixing distribution became computationally feasible either via an EM algorithm (as described in detail by McLachlan & Krishnan (1997) and McLachlan & Peel (2001) or via other algorithmic methods a detailed account of which is provided by Bohning (1995). Another example is the development of model-based clustering through mixture models. Such models became accessible by a wide range of research workers from a variety of disciplines, only after the systematic use of computers (see, e.g., McLachlan & Basford, 1989). Finally, Bayesian estimation for mixture models became possible via MCMC methods (see, e.g., Diebolt & Robert, 1994) that required high speed computer resources. The multiplicity of applications of mixtures presented in Section 3, reveals that the problems connected with the implementation of the theoretical results would not have become tractable if it were not for the advancement of computer technology. The results could have remained purely theoretical with a few applications by specialists in certain fields.

5. References

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical Modelling of Data on Teaching Styles. *Journal of the Royal Statistical Society*, *A* 144, 419-461.

Aitkin, M. and Wilson, T. (1980). Mixture Models, Outliers and the EM Algorithm. *Technometrics*, 22, 325-331.

Alanko, T. and Duffy, J.C. (1996). Compound Binomial Distributions for Modeling Consumption Data. *The Statistician*, 45, 269-286.

Al-Husainni E.K. and Abd-El-Hakim, N.S. (1989). Failure Rate of the Inverse Gaussian-Weibull Mixture Model. *Annals of the Institute of Statistical Mathematics,* 41, 617-622.

Andrews, D.F , Mallows, C. L. (1974). Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society*, *B* 36, 99-102.

Banfield, D.J. and Raftery, A.E. (1993). Model-based Gaussian and Non-Gaussian Clustering. *Biometrics,* 49, 803-821.

Barndorff-Nielsen, O.E. (1997). Normal Inverse Gaussian Distributions and Stochastic Volatility Modeling. *Scandinavian Journal of Statistics,* 24, 1-13.

Barndorff-Nielsen, O.E., Kent, J. and Sorensen, M. (1982). Normal Variance-Mean Mixtures and z-Distributions. *International Statistical Review*, 50, 145-159.

Bartholomew, D.J. (1980). Factor Analysis for Categorical Data. *Journal of the Royal Statistical Society*, B, 42, 292-321.

Böhning, D. (1995). A Review of Reliable Maximum Likelihood Algorithms for Semiparametric Mixture Models. *Journal of Statistical Planning and Inference*, 47, 5-28.

Böhning, D. (1999). *Computer Assisted Analysis of Mixtures (C.A.M.AN).* Marcel Dekker Inc. New York

Bose, S. (1994). Bayeasian Robustness with Mixture Classes of Priors. *Annals of Statistics*, 22, 652-667.

Brooks, R.J. (1984). Approximate Likelihood Ratio Tests in the Analysis of Beta-Binomial Data. *Applied Statistics*, 33, 285-289.

Brooks, S.P., Morgan, B.J.T., Ridout, M.S. and Pack, S.E. (1997). Finite Mixture Models for Proportions. *Biometrics*, 53, 1097-1115.

Cao, G. and West, M. (1996). Bayesian Analysis of Mixtures. *Biometrics*, 52, 221-227.

Caroll, R.J. and Hall, P. (1988). Optimal Rates of Convergence for Deconvoluting a Density. *Journal of the American Statistical Association*, 83, 1184-1186.

Chen, S.X. (2000). Probability Density Function Estimation Using Gamma Kernels. *Annals of the Institute of Statistical Mathematics,* 52, 471-490.

Dacey, M. F. (1972). A Family of Discrete Probability Distributions Defined by the Generalized Hyper-Geometric Series. *Sankhy* ā, B 34, 243-250.

Dalal, S. R. and Hall, W. J. (1983). Approximating Priors By Mixtures of Natural Conjugate Priors. *Journal of the Royal Statistical Society,* B 45, 278-286.

Dean, C.B., Lawless, J. and Willmot, G.E. (1989). A Mixed Poisson-Inverse Gaussian Regression Model. *Canadian Journal of Statistics,* 17, 171-182.

Devroye, L. (1992). *Non-Uniform Random Variate Generation*. Springer-Verlag. New York.

Diebolt, J. and Robert, C. (1994). Estimation of Finite Mixture Distributions Through Bayesian Sampling. *Journal of the Royal Statistical Society,* B 56, 363-375.

Eberlein, E. and Keller, U. (1995). Hyperbolic Distributions in Finance. *Bernoulli*, 1, 281-299.

Escobar, M. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90, 577-588.

Everitt, B.S. (1984). *An Introduction to Latent Variable Models.* Chapman and Hall. New York.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixtures Distributions*. Chapman and Hall. New York.

Everitt, B.S and Merette, C. (1990). The Clustering of Mixed-Mode Data: A Comparison of possible Approaches. *Journal of Applied Statistics,* 17, 283-297.

Gleser, L. J. (1989). The Gamma Distribution as a Mixture of Exponential Distributions. *American Statistician,* 43, 115-117.

Goldstein, H. (1995). *Multilevel Statistical Models.* (2nd edition) Arnold Publishers. London.

Goutis, K. (1993). Recovering Extra Binomial Variation*. Journal of Statistical Computation and Simulation,* 45, 233-242.

Greenwood, M. and Yule, G. (1920). An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurence of Multiple Attacks of Disease or of Repeated Accidents. *Journal of the Royal Statistical Society,* A 83, 255-279.

Gupta, S. and Huang, W.T. (1981). On Mixtures of Distributions: A Survey and Some New Results on Ranking and Selection. *Sankhya,* B 43, 245-290.

Hasselblad, V. (1969) Estimation of Finite Mixtures from the Exponential Family. *Journal of the American Statistical Association,* 64, 1459-1471.

Hebert, J. (1994). Generating Moments of Exponential Scale Mixtures. *Communications in Statistics- Theory and Methods*, 23, 1181-1189.

Hinde, J. and Demetrio, C. G. B. (1998). Overdispersion: Models and Estimation. *Computational Statistics and Data Analysis*, 27, 151-170.

Hosmer, D. (1973). A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of two Normal Distributions Under Three Different Types of Sample. *Biometrics,* 29, 761-770.

Irwin, J. O. (1963). The Place of Mathematics in Medical and Biological Statistics. *Journal of the Royal Statistical Society, A* 126, 1-44.

Irwin, J. O. (1968). The Generalized Waring Distribution Applied to Accident Theory. *Journal of the Royal Statistical Society, A* 131, 205-225.

Irwin, J. O. (1975). The Generalized Waring Distribution. *Journal of the Royal Statistical Society, A* 138, 18-31 (Part I), 204-227 (Part II), 374-384 (Part III).

Jewell, N. (1982). Mixtures of Exponential Distributions. *Annals of Statistics*, 10, 479-484.

Johnson, N.L., Kotz, S.and Kemp, A.W. (1992). *Univariate Discrete Distributions*. (2nd edition) Willey-New York.

Jorgensen, B., Seshadri V. and Whitmore G.A. (1991). On the Mixture of the Inverse Gaussian Distribution With its Complementary Reciprocal. *Scandinavian Journal of Statistics,* 18, 77-89.

Karlin, B. and Lewis, T. (1996). *Empirical Bayes Methods.* Chapman and Hall. New York.

Karlis, D. (1998). Estimation and Hypothesis Testing Problems in Finite Poisson Mixture. *Unpublished Ph.D. Thesis, Dept. of Statistics, Athens University of Economics and Business, ISBN 960-7929-19-5.*

Karlis, D. and Xekalaki, E. (1999). On Testing for the Number of Components in a Mixed Poisson Model. *Annals of the Institute of Statistical Mathematics,* 51, 149-162.

Kendall, M. G. (1961). Natural Law in the Social Sciences. *Journal of the Royal Statistical Society, A* 124, 1-16.

Krishnaji, N. (1970b). Characterization of the Pareto Distribution through a Model of Under-Reported Incomes. *Econometrica,* 38, 251-255.

Laird, N. (1982). Empirical Bayes Estimates Using the Nonparametric Maximum Likelihood Estimate for the Prior. *Journal of Statistical Computation and Simulation,* 15, 211-220.

Lawless, J. (1987). Negative Binomial and Mixed Poisson Regression. *Canadian Journal of Statistics,* 15, 209-225.

Lee, Y. and Nelder, J.A. (1996). Hierarchical Genaralized Linear Models. *Journal of the Royal Statistical Society,* B 58, 619-678.

Leroux, B (1992). Maximum Likelihood Estimation for Hidden Markov Models. *Stochastic Processes,* 49, 127-143.

Lindsay, B. (1983). The Geometry of Mixture Likelihood. A General Theory. *Annals of Statistics*, 11, 86-94.

Lindsay, B. (1995). *Mixture Models: Theory, Geometry and Applications*. Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics and American Statistical Association.

Lindsay, B., Clogg, C.C. and Grego, J. (1991). Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class for Item Analysis. *Journal of the American Statistical Association,* 86, 96-107.

Liu, M.C. and Taylor, R. (1989). A Consistent Nonparametric Density Estimator for the Deconvolution problem. *Canadian Journal of Statistics*, 17, 427-438.

Mallet, A. (1986). A Maximum Likelihood Estimation Method for Random Coefficient Regression Models. *Biometrika,* 73, 645-656.

Maritz, J. L. and Lwin, (1989). *Empirical Bayes* Methods. (2nd edition) Marcel Dekker Inc. New York.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience. New York.

McLachlan, G. and Basford, K. (1989). *Mixture Models: Inference and Application to Clustering*. Marcel Dekker Inc. New York.

McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and its Extensions*. Wiley. New York.

McLachlan, J.A. and Peel, D. (2001). *Finite Mixture Models*. Wiley. New York.

Mudholkar, G. and Chaubey, Y. P. (1976). On the Distribution of Fisher's Transformation of the Correlation Coefficient. *Communications in Statistics—Computation and Simulation,* 5, 163-172.

Panaretos, J. (1982). An Extension of the Damage Model. *Metrika*, 29, 189-194.

Panaretos, J. (1983). A Generating Model Involving Pascal and Logarithmic Series Distributions. *Communications in Statistics Part A: Theory and Methods*, Vol. A12, No.7, 841-848.

Philippe, A. (1997). Simulation of Right and Left Truncated Gamma Distributions by Mixtures. *Statistics and Computing*, 7, 173-181.

Rachev, St. and Sengupta, A. (1993). Laplace-Weibull Mixtures for Modeling Price Changes. *Management Science,* 39, 1029-1038.

Rao, C.R. (1963). On Discrete Distributions Arising out of Methods of Ascertainments. *Sankhya A,* 25, 311-324.

Redner R., Walker H. (1984). Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review,* 26, 195-230.

Rudas, T., Clogg, C.C. and Lindsay, B.G. (1994). A New Index of Fit Based on Mixture Methods for the Analysis of Contingency Tables. *Journal of the Royal Statistical Society,* B 56, 623-639.

Scallan, A. J. (1992). Maximum Likelihood Estimation for a Normal/Laplace Mixture Distribution. *The Statistician* 41, 227-231.

Shaked, M. (1980). On Mixtures from Exponential Families. *Journal of the Royal Statistical Society* B 42 , 192-198.

Sibuya, M. (1979). Generalised Hypergeometric, Digamma and Trigamma Distributions. *Annals of the Institute of Statistical Mathematics* A 31, 373-390.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. New York.

Simon, H. A. (1955). On a Class of Skew Distribution Functions. *Biometrika,* 42, 425-440.

Simonoff, J.S. (1996). *Smoothing Techniques.* Springer –Verlag. New York.

Symons, M., Grimson, R. and Yuan, Y. (1983). Clustering of Rare Events. *Biometrics* 39, 193-205.

Titterington, D.M. (1990). Some Recent Research in the Analysis of Mixture Distributions. *Statistics* 21, 619-641.

Titterington, D.M. Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixtures Distributions*. Wiley.

Tripathi, R., Gupta, R., Gurland, J. (1994). Estimation of Parameters in the Beta Binomial Model. *Annals of the Institute of Statistical Mathematics,* 46, 317-331.

Wang, M.C. and Van Ryzin, J. (1979). Discrete Density Smoothing Applied to the Empirical Bayes Estimation of a Poisson Mean. *Journal of Statistical Computation and Simulation*, 8, 207-226.

Wang, P., Puterman, M., Cokburn, I. and Le, N. (1996). Mixed Poisson Regression Models with Covariate Dependent Rates. *Biometrics*, 52, 381-400.

Wolfe, J.H. (1970). Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research,* 5, 329-350.

Xekalaki, E. (1983a). A Property of the Yule Distribution and its Applications. *Communications in Statistics, Part A, (Theory and Methods),* A12, 10, 1181-1189.

Xekalaki, E. (1983b). The Univariate Generalised Waring Distribution in Relation to Accident Theory: Proneness, Spells or Contagion? *Biometrics,* 39, 887-895.

Xekalaki, E. (1984a). The Bivariate Generalized Waring Distribution and its Application to Accident Theory. *Journal of the Royal Statistical Society,* A 147, 488-498.

Xekalaki, E. (1984b). Linear Regression and the Yule Distribution. *Journal of Econometrics,* 24 (1), 397-403.

Xekalaki, E. and Panaretos, J. (1983). Identifiability of Compound Poisson Distributions. *Scandinavian Actuarial Journal*, 66, 39-45.

Yule, G. U. (1925). A Mathematical Theory of Evolution Based on the Conclusions of Dr. J.G. Willis. *Philosophical Transactions of the Royal Society,* 213, 21-87.

Received: April 2002, Revised: June 2002