9

Finite Mixture Models with Nonnormal Components

9.1 Finite Mixtures of Exponential Distributions

9.1.1 Model Formulation and Parameter Estimation

It is often assumed that nonnegative observations are realizations of a random variable Y arising from a finite mixture of exponential distributions:

$$Y \sim \eta_1 \mathcal{E}(\lambda_1) + \cdots + \eta_K \mathcal{E}(\lambda_K), \tag{9.1}$$

where $\mathcal{E}(\lambda_k)$ is parameterized as in Appendix A.1.4. This mixture distribution is parameterized in terms of $\boldsymbol{\vartheta} = (\lambda_1, \ldots, \lambda_K, \boldsymbol{\eta})$. Teicher (1963) showed that mixtures of exponential distributions are identifiable.

Following Farewell (1982), various mixture survival models, based on the exponential or more general distributions, were suggested and studied by many authors; see, for instance, Morbiducci et al. (2003), who studied such models with special focus on cure-rate models, to estimate the unknown rate of cured patients and the survival function of uncured patients in a clinical trial. The popularity of these models in duration or survival analysis is explained by their ability to explain the frequently observed fact that hazards decline with the length of spells (Heckman et al., 1990). Another interesting application of mixtures of exponential distributions appears in failure analysis, where failure often occurs for more than one reason (Everitt and Hand, 1981; Taylor, 1995). Slud (1997) proposes a two-component exponential mixture model to test imperfect debugging in software reliability.

Heckman et al. (1990) consider a consistent method of moments estimator and present Bayesian and classical tests for testing the hypothesis of dealing with mixtures of exponential distributions. Taylor (1995) uses the EM algorithm. Gruet et al. (1999) use MCMC methods for Bayesian estimation and apply reversible jump MCMC to select the number of components.

9.1.2 Bayesian Inference

Bayesian estimation using data augmentation and MCMC as in *Algorithm 3.4* is easily implemented for a mixture of exponential distributions. Based on the prior $\lambda_k \sim \mathcal{G}(a_0, b_0)$, the complete-data posterior $p(\lambda_k|\mathbf{y}, \mathbf{S})$ is given by $\lambda_k|\mathbf{y}, \mathbf{S} \sim \mathcal{G}(a_k(\mathbf{S}), b_k(\mathbf{S}))$, where:

$$a_k(\mathbf{S}) = a_0 + N_k(\mathbf{S}),$$
$$b_k(\mathbf{S}) = b_0 + \sum_{i:S_i=k} y_i.$$

Gruet et al. (1999) show how a reparameterization of the exponential mixture model (9.1) can allow for noninformative priors. They count the mixture components starting from $k = 0$, rather than $k = 1$. They leave λ_0 and $\eta_0 = \omega_0$ unchanged, whereas each λ_k and η_k is expressed for $k = 1, \ldots, K-1$ as

$$\lambda_k = \lambda_0 \prod_{j=1}^{k} \tau_j,$$
$$\eta_k = (1-\omega_0)\cdots(1-\omega_{k-1})\omega_k.$$

This parameterization allows us to select a partially proper prior distribution, based on the improper $\mathcal{G}(0,0)$-prior for λ_0, whereas $\tau_1, \ldots, \tau_{K-1}$ are assumed to be uniform on [0,1]. As λ_0 appears as a common parameter in all component densities, this leads to a proper posterior density, as shown in the appendix of Gruet et al. (1999). This prior implies the order constraint $\lambda_0 > \cdots > \lambda_{K-1}$ on the component parameters, leading to an automatic identification of the model.

Casella et al. (2002) illustrate how perfect slice sampling may be implemented for mixtures of exponential distributions.

Reversible Jump MCMC

Gruet et al. (1999) apply reversible jump MCMC to select the number of components for an exponential mixture. Their parameterization introduces quite a natural strategy for carrying out split and merge moves, because in a mixture with $K-1$ components, the last component $\mathcal{E}(\lambda_0\tau_1\cdots\tau_{K-2})$ is replaced by a two-component exponential mixture:

$$\omega_{K-2}\mathcal{E}(\lambda_0\tau_1\cdots\tau_{K-2}) + (1-\omega_{K-2})\mathcal{E}(\lambda_0\tau_1\cdots\tau_{K-1})$$

to obtain a mixture with K components.

To perform a split move in a mixture with K components, first a component k is chosen randomly. The index of all components from $k+1, \ldots, K-1$ is shifted by one. To split the old component k into the two new components k and $k+1$, the new parameters τ_k^{new} and τ_{k+1}^{new} satisfy:

$$\tau_k^{new}\tau_{k+1}^{new} = \tau_k,$$

whereas the weights satisfy:

$$(1-\omega_k^{new})(1-\omega_{k+1}^{new}) = (1-\omega_k).$$

To perform the split move, two random numbers u_1 and u_2 are introduced:

$$\begin{aligned}
\tau_k^{new} &= u_1 + \tau_k(1-u_1),\\
\tau_{k+1}^{new} &= \frac{\tau_k}{\tau_k^{new}} = \frac{\tau_k}{u_1+\tau_k(1-u_1)},\\
\omega_k^{new} &= u_2\omega_k,\\
\omega_{k+1}^{new} &= \frac{\omega_k(1-u_2)}{1-\omega_k u_2}.
\end{aligned}$$

If $k > 0$, then $u_1, u_2 \sim \mathcal{U}[0,1]$, whereas $u_1 \sim \mathcal{U}[0,.5]$ for $k = 0$, in which case

$$\lambda_0^{new} = \lambda_0/u_1, \qquad \tau_1^{new} = u_1.$$

The determinant of the Jacobian is given by

$$|\text{Jacobian}| = \begin{cases} \dfrac{\omega_0(1-\tau_k)}{(1-\omega_k^{new})(\tau_k^{new})^2}, & \text{if } k > 0,\\[2ex] \dfrac{\omega_0}{(1-\omega_0^{new})u_1}, & \text{if } k = 0. \end{cases}$$

Gruet et al. (1999) report no improvement in refining reversible jumps by adding a move that introduces empty components.

9.2 Finite Mixtures of Poisson Distributions

9.2.1 Model Formulation and Estimation

A popular model for describing the distribution of count data is the Poisson mixture model, where it is assumed that $y_1, \ldots, y_N$ are independent realization of a random variable Y arising from a mixture of Poisson distributions:

$$Y \sim \eta_1\mathcal{P}(\mu_1) + \cdots + \eta_K\mathcal{P}(\mu_K),$$

with $\mathcal{P}(\mu_k)$ being a Poisson distribution with mean μ_k; see Appendix A.1.11. This distribution is parameterized in terms of $2K-1$ distinct model parameters $\boldsymbol{\vartheta} = (\mu_1, \ldots, \mu_K, \boldsymbol{\eta})$. Mixtures of Poisson distributions are identifiable; see Feller (1943) and Teicher (1960).

Applications of mixtures of Poisson distributions appear in particular in biology and medicine; see, for example, Farewell and Sprott (1988) and Pauler

et al. (1996). Applications to disease mapping are briefly discussed in Subsection 9.4.1. Karlis and Xekalaki (2005) provide a recent review of Poisson mixtures.

Mixtures of Poisson distributions served to illustrate statistical inference for finite mixtures throughout Chapter 2 to Chapter 5. Reversible jump MCMC has been used for finite mixtures of Poisson distributions by Dellaportas et al. (2002) and Viallefont et al. (2002); see also Subsection 5.2.2.

For Bayesian estimation, we add only comments on choosing the hyperparameters a_0 and b_0 of the prior of the group means, $\mu_k \sim \mathcal{G}(a_0, b_0)$. Viallefont et al. (2002) suggest fixing a_0 around 1 and choosing b_0 in such a way that the prior mean $\mathrm{E}(Y|\boldsymbol{\vartheta}) = a_0/b_0$ is matched to the midrange of the data, for example, the mean:

$$b_0 = \frac{a_0}{\overline{y}}. \tag{9.2}$$

For data where overdispersion is actually present, meaning that $s_y^2 - \overline{y} > 0$, it is possible to choose a_0 in such a way that the expectation of the second factorial moment with respect to the $\mathcal{G}(a_0, b_0)$-prior, which is by $\mathrm{E}(Y(Y-1)|\boldsymbol{\vartheta}) = a_0/b_0^2(1 + 1/a_0)$, is matched to the second factorial moment of the data, v_2, defined earlier in (2.26):

$$a_0 = \frac{\overline{y}^2}{v_2 - \overline{y}^2}, \tag{9.3}$$

where due to (9.6) $v_2 - \overline{y}^2$ could be substituted by $s_y^2 - \overline{y}$. Thus the larger the overdispersion in the data is, the smaller a_0 should be chosen.

If overdispersion is small, then a_0 is large and μ_k is strongly shrunken toward a_0/b_0. In this case it useful to assume a hierarchical prior as defined in Subsection 3.2.4, where $b_0 \sim \mathcal{G}(g_0, G_0)$. Estimation and model selection are rather insensitive to the parameter g_0 and could be chosen as $g_0 = 0.5$, whereas matching $\mathrm{E}(b_0) = g_0/G_0$ to $a_0/\overline{y}$ yields:

$$G_0 = \frac{g_0 \overline{y}}{a_0}.$$

9.2.2 Capturing Overdispersion in Count Data

Overdispersion occurs for a random variable Y, if the variance is bigger than the mean, whereas mean and variance are identical for a Poisson distribution. Overdispersion is present in many data sets involving counts. For illustration, consider the Eye Tracking Data counting eye anomalies in 101 schizophrenic patients studied by Pauler et al. (1996) and Escobar and West (1998), where the sample variance $s_y^2 = 35.89$ shows overdispersion in comparison to the sample mean $\overline{y} = 3.5248$; see also the histogram of the data in Figure 9.1.

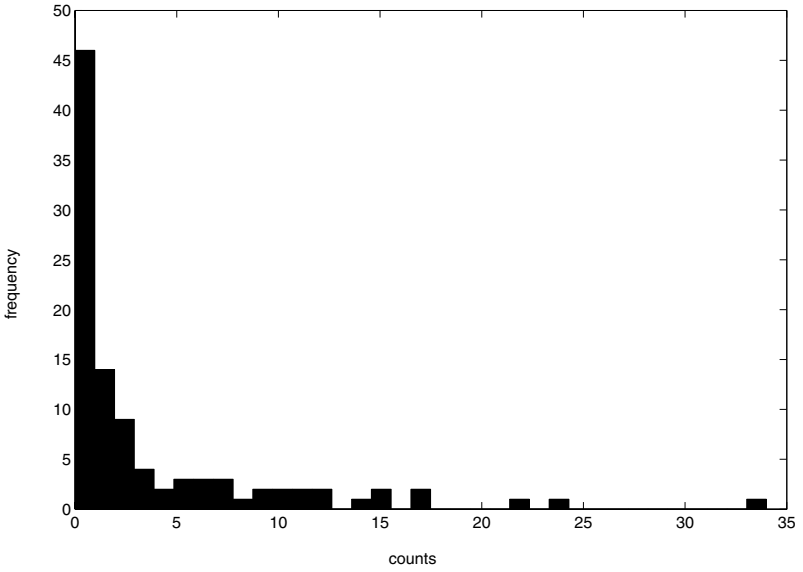

Fig. 9.1. Eye Tracking Data, empirical distribution of the observations

Many authors have studied the effect of overdispersion; see Wang et al. (1996) for some review. One possible reason for overdispersion is unobserved heterogeneity in the sample, causing the mean to be different among the observed subjects. A model commonly used in this context and discussed already in Feller (1943), is the Poisson–Gamma model which is a continuous mixture of Poisson distributions:

$$Y \sim \mathcal{P}\left(\mu_i^s\right), \qquad \mu_i^s \sim \mathcal{G}\left(\alpha, \alpha/\mu\right). \tag{9.4}$$

Marginally, Y arises from the NegBin $(\alpha, \alpha/\mu)$-distribution, with $\mathrm{E}(Y|\boldsymbol{\vartheta}) = \mu$ and

$$\mathrm{Var}(Y|\boldsymbol{\vartheta}) = \mathrm{E}(Y|\boldsymbol{\vartheta})\frac{\alpha + \mathrm{E}(Y|\boldsymbol{\vartheta})}{\alpha} \geq \mathrm{E}(Y|\boldsymbol{\vartheta}),$$

where $\boldsymbol{\vartheta} = (\alpha, \mu)$. As long as α is not too large, this distribution actually captures overdispersion.

Overdispersion of a random variable Y drawn from a Poisson mixture is evident from the first two moments of this mixture given by (1.19):

$$\mathrm{E}(Y|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \mu_k \eta_k,$$

$$\mathrm{Var}(Y|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \mu_k(1+\mu_k)\eta_k - \mathrm{E}(Y|\boldsymbol{\vartheta})^2 = \mathrm{E}(Y|\boldsymbol{\vartheta}) + B(\boldsymbol{\vartheta}),$$

where $B(\boldsymbol{\vartheta})$ is the between-group heterogeneity:

$$B(\boldsymbol{\vartheta}) = \sum_{k=1}^{K} (\mu_k - \mu(\boldsymbol{\vartheta}))^2 \eta_k, \tag{9.5}$$

with $\mu(\boldsymbol{\vartheta}) = \mathrm{E}(Y|\boldsymbol{\vartheta})$. As $\mathrm{Var}(Y|\boldsymbol{\vartheta}) - \mathrm{E}(Y|\boldsymbol{\vartheta}) = B(\boldsymbol{\vartheta})$, finite mixtures of Poisson distributions explain overdispersion through unobserved heterogeneity in the sample, causing the mean to be different among the observed subjects. For $K = 2$, for instance, $B(\boldsymbol{\vartheta}) = 2\eta_1\eta_2(\mu_2 - \mu_1)^2$. Overdispersion occurs as long as the means of at least two components are different. Overdispersion could also be determined from the difference of the second factorial moment of the Poisson mixture, $\mathrm{E}(Y(Y-1)|\boldsymbol{\vartheta})$, and $\mathrm{E}(Y|\boldsymbol{\vartheta})^2$, as $\mathrm{E}(Y(Y-1)|\boldsymbol{\vartheta}) = \mathrm{E}(Y^2|\boldsymbol{\vartheta}) - \mathrm{E}(Y|\boldsymbol{\vartheta})$, and therefore:

$$\mathrm{E}(Y(Y-1)|\boldsymbol{\vartheta}) - \mathrm{E}(Y|\boldsymbol{\vartheta})^2 = B(\boldsymbol{\vartheta}). \tag{9.6}$$

The use of finite mixtures of Poisson distributions, rather than the more commonly used Poisson–Gamma model, to account for overdispersion has attracted several researchers, among them Simar (1976), Manton et al. (1981), Lawless (1987), Leroux (1992a), Leroux and Puterman (1992), Wang et al. (1996), and Viallefont et al. (2002).

It is possible to include observed covariates to explain part of the unobserved heterogeneity as discussed in Subsection 9.4.1, dealing with mixtures of Poisson regression models.

9.2.3 Modeling Excess Zeros

Count data often contain more zeros than expected under the Poisson distribution. In medical data excess zeros occur if the zero-class is inflated by the inclusion of observations that belong to a noninfected group. The Eye Tracking Data, for instance, contain 46 zeros, whereas under the $\mathcal{P}(\mu)$-distribution, the number of zeros in a sample of size N follows a BiNom$(N, e^{-\mu})$-distribution. For $N = 101$ and $\mu = \overline{y} = 3.5248$, the expected number of zero counts is roughly equal to 3, whereas the probability to observe at least 46 zero counts in a sample from the $\mathcal{P}(3.5248)$-distribution is as small as $1.9 \cdot 10^{-14}$, clearly indicating the presence of excess zeros.

Analyzing count data with excess zeros, sometimes also called inflated zeros, has a long tradition in applied statistics; see Meng (1997) for an interesting review. Feller (1943) proves that the number of zeros in a Poisson mixture is always larger than the number of zeros in a single Poisson distribution $\mathcal{P}(\mu)$ with the same mean $\mu = \mathrm{E}(Y|\boldsymbol{\vartheta})$ as the mixture distribution. This follows immediately from:

$$\Pr(Y=0|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k e^{-\mu_k} = e^{-\mu} \sum_{k=1}^{K} \eta_k e^{\mu-\mu_k} \tag{9.7}$$
$$\geq e^{-\mu} \sum_{k=1}^{K} \eta_k (1+\mu-\mu_k) \geq e^{-\mu}.$$

Cohen (1966) considers the following two-component mixture:

$$Y \sim \eta_1 I_{\{0\}}(y_i) + \eta_2 \mathcal{P}(\mu_2), \tag{9.8}$$

where $I_{\{0\}}(y_i)$ is 1 iff $y_i = 0$. A limitation of (9.8) is that the second group is assumed to be homogeneous. To capture overdispersion among nonzero individuals, it is sensible to substitute the Poisson distribution by a more general distribution, such as a finite mixture of $K-1$ Poisson distributions as in Cohen (1960) or a negative binomial distribution as in Cohen (1966). Such models are known as hurdle models; see, for instance, Cameron and Trivedi (1998) and Dalrymple et al. (2003) for an application to sudden infant death syndrome.

9.2.4 Application to the Eye Tracking Data

For illustration, consider the count data on eye tracking anomalies in 101 schizophrenic patients studied by Escobar and West (1998). To capture overdispersion and excess zeros for this data set, diagnosed in Subsection 9.2.2, we model the data by a finite mixture of K Poisson distributions as in Congdon (2001), with increasing number K of potential groups. We use the hierarchical prior (3.12) with $a_0 = 0.1$, $g_0 = 0.5$, and $G_0 = g_0\overline{y}/a_0$, and a $\mathcal{D}(4,\ldots,4)$-prior for $\boldsymbol{\eta}$. We use *Algorithm 3.3* for MCMC estimation, and store 8000 MCMC draws after a burn-in-phase of 3000.

Figure 9.2 shows, for an increasing number of components $K = 1,\ldots,7$, the posterior distribution of the probability $p_0(\boldsymbol{\vartheta})$ to observe 0, which is given by (9.7), of the overdispersion $B(\boldsymbol{\vartheta})$ defined in (9.5), and of the lth factorial moment, $\sum_{k=1}^{K} \eta_k \mu_k^l$ for $l = 3$ and $l = 4$. A comparison of these posterior distributions to the corresponding sample moments indicates that either four or five components are sufficient to capture the moments under investigation. Adding additional components hardly changes the posterior distribution of these moments.

Formal model selection, either using marginal likelihoods or reversible jump MCMC, is not really conclusive. Table 9.1 shows the log of the marginal likelihood for an increasing number of components, estimated through various simulation-based approximations, that were discussed in Section 5.4, namely bridge sampling, importance sampling, and reciprocal importance sampling. The importance density is constructed from the MCMC draws as in (5.36) with $S = \min(50K!, 5000)$, and the estimators are based on $M = 5000$ MCMC draws and $L = 5000$ draws from the importance density. Up to $K = 4$, these

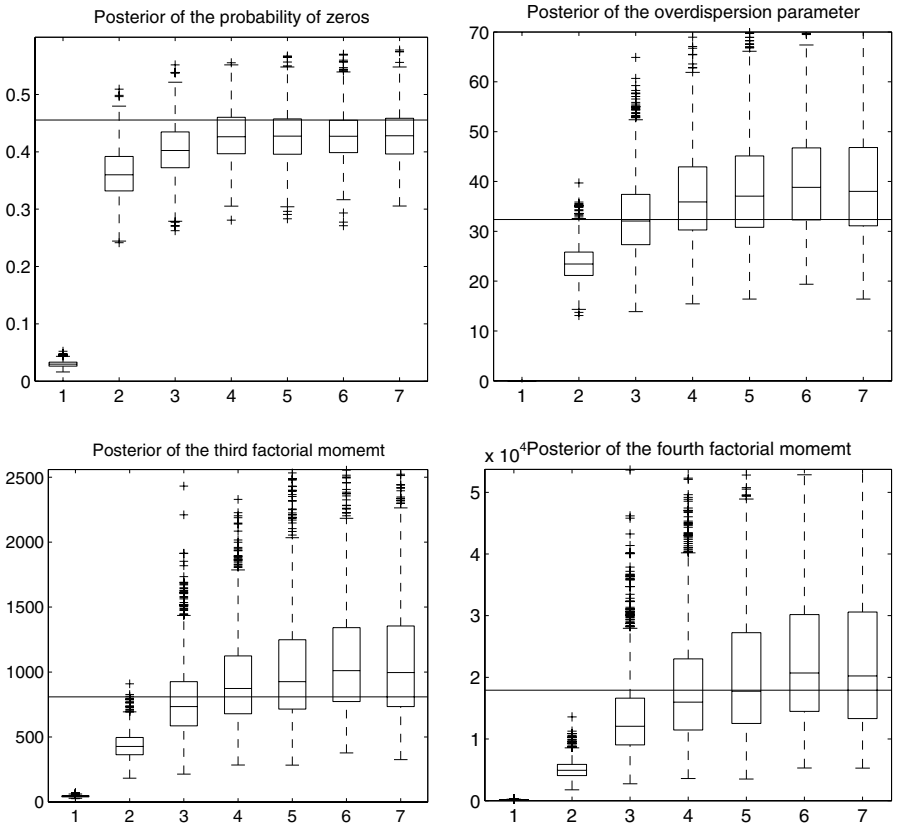

Fig. 9.2. Eye Tracking Data, finite Poisson mixtures with increasing numbers K of potential groups; posterior distribution of the probability $p_0(\boldsymbol{\vartheta})$ to observe 0 (top left), of the overdispersion $B(\boldsymbol{\vartheta})$ (top right), the third (bottom left), and the fourth (bottom right) factorial moment in comparison to the corresponding sample moments (black horizontal line) for $K = 1, \ldots, 7$

Table 9.1. Eye Tracking Data, various estimators of the marginal likelihood $p(\mathbf{y}|K)$ for finite mixtures of Poisson distributions with $K = 1$ to $K = 7$ components

$p(\mathbf{y}\|K)$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
$\log \hat{p}_{BS}(\mathbf{y}\|K)$	–472.02	–252.61	–237.29	–232.81	–232.55	–234.07	–235.68
$\log \hat{p}_{IS}(\mathbf{y}\|K)$	–472.02	–252.62	–237.28	–232.67	–231.08	–230.37	–231.53
$\log \hat{p}_{RI}(\mathbf{y}\|K)$	–472.02	–252.61	–237.32	–233.40	–234.44	–236.74	–238.28

estimators are rather similar, but from $K = 5$ onwards, the estimators tend to be rather unstable. This leads to quite different estimators for the model posterior probabilities $\Pr(K|\mathbf{y})$, displayed in Table 9.2, which were computed under the truncated Poisson prior $p(K) \propto f_P(K;1)$. Although all estimators suggest choosing $K = 4$, the estimated posterior probabilities are quite different, and differ substantially from the posterior probabilities obtained from reversible jump MCMC, which are given in the same table. By considering a different importance density, namely a full permutation of single draw $\mathbf{S}^\star$, we obtained estimators of the model probabilities that are rather close to the estimators obtained from reversible jump MCMC. This suggests that simulation-based approximations to the marginal likelihood are sensitive for K larger than 3 or 4, and reversible jump MCMC is preferable for mixtures with a medium to large number of components.

The same table shows that AIC and BIC also lead to the conclusion to choose $K = 4$, however, again evidence in favor of this model is very fragile, as AIC for $K = 4$ is only slightly larger than AIC for $K = 5$, whereas BIC for $K = 4$ is only slightly larger than BIC for $K = 3$.

Table 9.2. Eye Tracking Data, posterior probabilities $\Pr(K|\mathbf{y})$ based on the prior $p(K) \propto f_P(K;1)$, obtained from $\log \hat{p}_{BS}(\mathbf{y}|K)$ and reversible jump MCMC (RJMCM), AIC and BIC for $K = 1$ to $K = 7$ number of components

$\Pr(K\|\mathbf{y})$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
Based on $\hat{p}_{BS}(\mathbf{y}\|K)$	0.00	0.00	0.03	0.76	0.20	0.01	0.00
Based on $\hat{p}_{IS}(\mathbf{y}\|K)$	0.00	0.00	0.02	0.42	0.41	0.14	0.01
Based on $\hat{p}_{RI}(\mathbf{y}\|K)$	0.00	0.00	0.07	0.87	0.06	0.00	0.00
RJMCMC	0.00	0.00	0.01	0.33	0.40	0.20	0.06
Based on $\hat{p}_{BS,2}(\mathbf{y}\|K)$	0.00	0.00	0.02	0.36	0.32	0.22	0.09
AIC	–472.02	–247.48	–230.22	–227.60	–227.94	–229.94	–231.94
BIC	–472.02	–251.40	–236.76	–236.76	–239.71	–244.32	–248.94

To obtain estimates of the group means and group sizes for a mixture of $K = 4$ groups, we need to identify a unique labeling among the MCMC draws. We first ran Gibbs sampling for an unconstrained mixture model with $K = 4$, and found that label switching took place between the two groups with the smallest means. For this data set, we could not achieve a unique labeling through unsupervised clustering. Next we imposed the constraint $\mu_1 < \cdots < \mu_K$ using the permutation sampler. Whenever the constraint was violated, we reordered the MCMC output in such a way that the constraint is fulfilled. Imposing the constraint eliminated label switching. Table 9.3 summarizes estimates of all group means and group sizes for $K = 4$, respectively. Evidently, the first group is practically a zero-movement group.

Table 9.3. EYE TRACKING DATA, posterior inference based on a Poisson mixture with $K = 4$ groups (hierarchical prior with $a_0 = 0.1$, $g_0 = 0.5$, and $G_0 = g_0\overline{y}/a_0$); identification obtained through imposing the constraint $\mu_1 < \cdots < \mu_4$

k	$\mathrm{E}(\mu_k\|\mathbf{y})$	(95% Confidence Region)	$\mathrm{E}(\eta_k\|\mathbf{y})$	(95% Confidence Region)
1	0.09	(0.00,0.38)	0.36	(0.19,0.55)
2	1.38	(0.61,2.89)	0.33	(0.17,0.49)
3	7.95	(5.07,10.83)	0.20	(0.12,0.30)
4	20.17	(15.16,27.55)	0.10	(0.04,0.18)

9.3 Finite Mixture Models for Binary and Categorical Data

9.3.1 Finite Mixtures of Binomial Distributions

For binomial mixtures the component densities arise from $\mathrm{BiNom}\,(T, \pi)$-distributions, where T is commonly assumed to be known, whereas the component-specific probabilities π are unknown and heterogeneous:

$$Y \sim \eta_1 \mathrm{BiNom}\,(T, \pi_1) + \cdots + \eta_K \mathrm{BiNom}\,(T, \pi_K)\,.$$

The density of this mixture is given by

$$p(y|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \binom{T}{y} \pi_k^y (1-\pi_k)^{T-y}, \tag{9.9}$$

with $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_K, \eta_1, \ldots, \eta_K)$. Binomial mixtures are not necessarily identifiable, as discussed already in Subsection 1.3.4. A necessary and sufficient condition is $T \geq 2K - 1$; see Teicher (1961).

Ever since Pearson (1915) employed a mixture of two binomial distributions to model yeast cell count data, discrete as well as continuous binomial mixtures have been suggested as overdispersed alternatives to the binomial distribution. Farewell and Sprott (1988), for instance, discuss an application in medicine to model the effect of a drug on patients who experience frequent premature ventricular contraction, whereas Brooks et al. (1997) and Brooks (2001) apply finite mixtures of binomials to dominant-lethal testing in a biological experiment.

Finite mixtures of binomial distributions may be extended to the case where T_i varies between the realizations $y_1, \ldots, y_N$:

$$p(y_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \binom{T_i}{y_i} \pi_k^{y_i} (1-\pi_k)^{T_i - y_i}.$$

For identifiability of the corresponding mixture we refer to Teicher (1963, p.1268). Böhning et al. (1998) discuss an application of this model to a prevalence study in veterinary science.

Unobserved Heterogeneity in Occurrence Probabilities

Assume that K hidden groups are present in a population with heterogeneity in the occurrence probability of a certain event, such as the choice probability for a certain product. Let $\pi_1, \ldots, \pi_K$ denote the different probabilities. Assume that for N randomly selected subjects it is observed if the event under investigation has occurred or not, with $Y_i = 1$ indicating occurrence.

The identifiability condition discussed above becomes essential if we want to use this information to estimate the unknown probabilities $\pi_1, \ldots, \pi_K$ as well as the unknown group sizes $\eta_1, \ldots, \eta_K$. Evidently $\Pr(Y_i = 1|\boldsymbol{\vartheta}) = \pi_k$, if subject i belongs to class k. As we observed for each subject only once whether the event has occurred, the marginal distribution of Y_i is a mixture of binomial distributions with $T = 1$, which is not identified. Hence it is not possible to estimate $\boldsymbol{\vartheta}$.

To this aim, it is necessary to observe the event under investigation more than once, thus we need repeated measurements $Y_{it}, t = 1, \ldots, T$ for each subject. In this case, the distribution of the number of successes, $Y_i = \sum_{t=1}^{T} Y_{it}$, is a realization from a mixture of binomial distributions $\text{BiNom}\,(T, \pi_k)$, which is identifiable only if $T \geq 2K - 1$. Thus even two repeated measurements are not sufficient to estimate the unknown parameter $\boldsymbol{\vartheta}$. For $K = 2$, for instance, the identifiability condition implies that we need for each subject i at least $T = 3$ repeated measurements on occurrence/nonoccurrence of the event in order to identify the group sizes and the probabilities. With increasing number of hidden groups, the number of repeated measurement increases.

It is possible to include observed covariates to explain part of the heterogeneity in the occurrence probabilities $\pi_1, \ldots, \pi_K$, as discussed in Subsection 9.4.2.

Extra-Binomial Variation

Extra-binomial variation, meaning that $\text{Var}(Y|\boldsymbol{\vartheta}) > \text{E}(Y|\boldsymbol{\vartheta})(1 - \text{E}(Y|\boldsymbol{\vartheta})/T)$, is present in many data sets involving binary data. Extra-binomial variation is often due to unobserved heterogeneity in the population, for example, if an important covariate is omitted.

A common way of dealing with extra-binomial variation is the Beta-binomial model, which is a continuous mixture of binomial distributions, where $Y \sim \text{BiNom}\,(T, \pi_i^s)$ and $\pi_i^s \sim \mathcal{B}\,(\alpha, \beta)$. Marginally, this leads to the Beta-binomial distribution:

$$p(y|\boldsymbol{\vartheta}) = \binom{T}{y} \frac{B(\alpha + y, \beta + T - y)}{B(\alpha, \beta)},$$

with $\boldsymbol{\vartheta} = (\alpha, \beta)$. The first two moments of this distribution read with $\pi = \alpha/\beta$:

$$\begin{aligned} \text{E}(Y|\boldsymbol{\vartheta}) &= T\pi, \\ \text{Var}(Y|\boldsymbol{\vartheta}) &= T\pi(1 - \pi) + (T - 1)T\frac{\pi(1 - \pi)}{\alpha + \beta + 1}. \end{aligned}$$

A finite mixture of binomial distributions is an interesting alternative to the Beta-binomial distribution. Overdispersion of a random variable Y, drawn from the binomial mixture (9.9) is evident from the first two moments of this mixture, which are easily derived from (1.19):

$$\begin{aligned} \mathrm{E}(Y|\boldsymbol{\vartheta}) &= T\pi, \qquad \pi = \sum_{k=1}^{K} \eta_k \pi_k, \\ \mathrm{Var}(Y|\boldsymbol{\vartheta}) &= T\pi(1-\pi) + (T-1)T\left(\sum_{k=1}^{K} \eta_k \pi_k^2 - \pi^2\right). \end{aligned} \tag{9.10}$$

For $T > 1$ extra variation due to the second term in (9.10) is present for any mixture with at least two different occurrence probabilities.

Bayesian Estimation of Binomial Finite Mixture Models

Bayesian inference for mixtures of binomial distributions is considered in Brooks (2001), who applied a Metropolis–Hastings algorithm. Bayesian estimation using data augmentation and MCMC as in *Algorithm 3.4* is easily implemented for a mixture of binomial distributions; see again Brooks (2001). Based on the conjugate Beta prior $\pi_k \sim \mathcal{B}(a_0, b_0)$, the posterior $p(\pi_k|\mathbf{y}, \mathbf{S})$ is again a Beta distribution, $\pi_k|\mathbf{y}, \mathbf{S} \sim \mathcal{B}(a_k(\mathbf{S}), b_k(\mathbf{S}))$, where:

$$\begin{aligned} a_k(\mathbf{S}) &= a_0 + \sum_{i:S_i=k} y_i, \\ b_k(\mathbf{S}) &= b_0 + \sum_{i:S_i=k} (T - y_i). \end{aligned}$$

Brooks (2001) applies the reversible jump MCMC method to jump between mixtures with different number of components and between mixtures of binomial distributions and mixtures of Beta-binomial distributions, where the number of components is left unchanged.

9.3.2 Finite Mixtures of Multinomial Distributions

Consider a categorical variable of more than two categories $\{1, \ldots, D\}$. Let Y_l, for $l = 1, \ldots, D$, be the number of occurrences of category l among T trials. If the occurrence probability distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)$ of each category is homogeneous among the observed subjects, then $\mathbf{Y} = (Y_1, \ldots, Y_D) \sim \mathrm{MulNom}(T, \boldsymbol{\pi})$; see also Appendix A.1.8 for a definition of the multinomial distribution.

To deal with unobserved heterogeneity in the occurrence probability of the various categories, $\mathbf{Y} = (Y_1, \ldots, Y_D)$ is assumed to follow a finite mixture of multinomial distributions,

$$Y \sim \eta_1 \text{MulNom}\,(T, \pi_1) + \cdots + \eta_K \text{MulNom}\,(T, \pi_K)\,,$$

with $\boldsymbol{\pi}_k = (\pi_{k,1}, \ldots, \pi_{k,D})$ being the unknown occurrence probability in group k. The density is given by

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \binom{T}{y_1 \ldots y_D} \prod_{l=1}^{D} \pi_{k,l}^{y_l}, \tag{9.11}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K, \boldsymbol{\eta})$,

Morel and Nagaraj (1993) use such a model for capturing multinomial extra variation. In this respect, the finite mixture distribution (9.11) is an interesting alternative to the more commonly applied Dirichlet-multinomial distribution, where $\mathbf{Y} \sim \text{MulNom}\,(T, \boldsymbol{\pi}^s)$ and $\boldsymbol{\pi}^s \sim \mathcal{D}\,(\alpha_1, \ldots, \alpha_D)$; see, for instance, Paul et al. (1989) and Kim and Margolin (1992).

Further applications are found in clustering Internet traffic (Jorgensen, 2004) and developmental psychology (Cruz-Medina et al., 2004). Banjeree and Paul (1999) extend (9.11) to deal with multinomial clustered data. Further extensions are finite mixtures of multinomial logit models that are discussed in Subsection 9.4.1.

Bayesian Estimation of Multinomial Finite Mixture Models

Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{iD})$, $i = 1, \ldots, N$, be N observations, where each $y_{il}, l = 1, \ldots, D$, counts the number of occurrences of category l in a series of T independent Bernoulli trials. Assume that a finite mixture of multinomial distributions should be fitted to these data.

Bayesian estimation using data augmentation and MCMC as in *Algorithm 3.4* is easily implemented for a mixture of multinomial distributions. Let $\boldsymbol{\pi}_k = (\pi_{k,1}, \ldots, \pi_{k,D})$ be the unknown discrete probability distribution in group k. Based on the Dirichlet prior $\boldsymbol{\pi}_k \sim \mathcal{D}\,(a_{0,1}, \ldots, a_{0,D})$, the posterior $p(\boldsymbol{\pi}_k|\mathbf{y}, \mathbf{S})$ is again a Dirichlet distribution $\boldsymbol{\pi}_k|\mathbf{y}, \mathbf{S} \sim \mathcal{D}\,(a_{k,1}(\mathbf{S}), \ldots, a_{k,D}(\mathbf{S}))$, where:

$$a_{k,l}(\mathbf{S}) = a_{0,l} + \sum_{i:S_i=k} y_{il}, \qquad l = 1, \ldots, D.$$

9.4 Finite Mixtures of Generalized Linear Models

Any of the finite mixture models discussed earlier in this chapter may be extended by assuming that in each group the underlying discrete distribution depends on some covariates. A common way to accommodate dependence of a nonnormal distribution on covariates is the generalized linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1999). Finite mixtures of generalized linear models extend the finite mixture of regression models discussed

in Chapter 8 to nonnormal data and find numerous applications in particular in biology, medicine, and marketing. A very useful taxonomical review of numerous applications of mixture regression models for various types of data may be found in Wedel and DeSarbo (1993b, Table 10.1), which includes a lot of additional references.

After reviewing in Subsections 9.4.1 and 9.4.2 some specific examples for count, binary, and multinomial data, estimation of such models is discussed in detail in Subsection 9.4.3.

9.4.1 Finite Mixture Regression Models for Count Data

Finite mixture regression models for count data are either based on the Poisson or the negative binomial distribution.

Finite Mixtures of Poisson Regression Models

Let Y_i denote the ith response variable, observed in reaction to a covariate $\mathbf{x}_i$, where the last element of $\mathbf{x}_i$ is 1, corresponding to an intercept. It is assumed that the marginal distribution of Y_i follows a mixture of Poisson distributions,

$$Y_i \sim \sum_{k=1}^{K} \eta_k \mathcal{P}\left(\mu_{k,i}\right), \tag{9.12}$$

where $\mu_{k,i} = \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$. If exposure data e_i are available for each subject, then $\mu_{k,i} = e_i \exp(\mathbf{x}_i \boldsymbol{\beta}_k)$. If $\mathbf{x}_i = 1$, a finite mixture of Poisson distributions with $\mu_k = \exp(\boldsymbol{\beta}_k)$ results; if $K = 1$, the standard Poisson regression model results.

For a standard Poisson regression model, conditional on a given covariate, the data often exhibit overdispersion. Wang et al. (1996) showed that a mixture of Poisson regression models is able to capture overdispersion. For a fixed covariate, the mean and variance of Y_i are easily obtained as in Subsection 9.2.2:

$$\begin{aligned}
\mathrm{E}(Y_i|\boldsymbol{\vartheta}) &= \sum_{k=1}^{K} \eta_k \mu_{k,i}, \\
\mathrm{Var}(Y_i|\boldsymbol{\vartheta}) &= \mathrm{E}(Y_i|\boldsymbol{\vartheta}) + \left(\sum_{k=1}^{K} \eta_k \mu_{k,i}^2 - \mathrm{E}(Y_i|\boldsymbol{\vartheta})^2 \right).
\end{aligned}$$

Wang et al. (1996) falsely claim that a mixture of Poisson regression models is identifiable if the regressor matrix is of full rank. However, as discussed in Subsection 8.2.2 for mixtures of normal regression models, this condition is in general not sufficient.

Pointwise identifiability for a fixed covariate $\mathbf{x}_i$ follows from the generic identifiability of Poisson mixtures. Thus

$$\sum_{k=1}^{K} \eta_k f_P(y; \mu_{k,i}) = \sum_{k=1}^{K} \eta_k^\star f_P(y; \mu_{k,i}^\star),$$

where $\log \mu_{k,i} = \mathbf{x}_i \boldsymbol{\beta}_k$ and $\log \mu_{k,i}^\star = \mathbf{x}_i \boldsymbol{\beta}_k^\star$, implies that the condition

$$\eta_k^\star = \eta_{\rho_i(k)}, \qquad \mathbf{x}_i \boldsymbol{\beta}_k^\star = \mathbf{x}_i \boldsymbol{\beta}_{\rho_i(k)}, \tag{9.13}$$

holds for some permutation $\rho_i(\cdot)$. This is exactly what resulted for finite mixtures of the standard regression model studied earlier in Section 8.2.2. It follows immediately that a mixture of Poisson regressions, where only the intercept is switching, is identifiable. In all other cases, evidently the same conditions as for a Gaussian mixture of regression models hold.

Many applications of this model appear in medicine such as modeling epileptic seizure frequency data in a clinical trial (Wang et al., 1996; Wang and Puterman, 1999) or modeling the length of hospital stay (Lu et al., 2003). Wedel et al. (1993) and Wedel and DeSarbo (1995) discuss applications in marketing, such as modeling the number of coupons used by a household and evaluating direct marketing strategies. Applications in road safety appear in Viallefont et al. (2002) and Hurn et al. (2003), who relate the number of accidents to covariates.

Disease Mapping

An area where mixtures of Poisson regression models are applied frequently is the study of disease distributions. The analysis of the geographic variation of disease and the representation of a disease distribution on a map is one of the oldest applications of statistics in epidemiology; see Schlattmann and Böhning (1993) for a review. Simple probabilistic models are based on the assumption that the number Y_i of cases observed in region i follows a $\mathcal{P}(\lambda e_i)$-distribution, where λ is the relative risk and e_i are the exposures. Rather than assuming that the risk is the same in all areas, Schlattmann and Böhning (1993) consider the case where the relative risk differs among the different areas, and takes one out of K values $\lambda_1, \ldots, \lambda_K$; see also Viallefont et al. (2002). This model is extended in Schlattmann et al. (1996) to accommodate dependence of covariates $\mathbf{x}_i = \begin{pmatrix} z_{i1} & \cdots & z_{i,d} \end{pmatrix}$ measured in each area:

$$Y_i | S_i \sim \mathcal{P}\left(\lambda_{S_i} \exp(\mathbf{x}_i \boldsymbol{\beta}) e_i\right),$$

whereas Viallefont et al. (2002) also consider heterogeneous covariate effects $\boldsymbol{\beta}_{S_i}$. Marginally, these models are finite mixtures of regression models that allow the detection of disease clusters, that is, areas of high or low risk. It provides an alternative to hierarchical Bayesian models for disease mapping; see, for instance, Bernardinelli et al. (1995).

Extensions of these models which substitute the unrealistic independence assumption among the indicators $S_1, \ldots, S_N$ by a spatial dependence model are discussed, among others, by Fernández and Green (2002) and Green and Richardson (2002).

Zero-Inflated Poisson Regressions

Lambert (1992) proposed the zero-inflated Poisson mixture regression model for dealing with zero-inflated count data with covariates and discussed an application where a production system switches between a perfect state where defects are extremely rare and an imperfect state where defects are possible. Both η_1, the probability of the perfect state as well as μ_2, the mean of the imperfect state depend on covariates through a logit-type model. Further applications are disease mapping (Böhning, 1998) and the analysis of sudden infant death syndrome in relation to climate (Dalrymple et al., 2003).

Finite Mixtures of Negative Binomial Regression Models

Ramaswamy et al. (1994) apply a finite mixture of negative binomial regression models in marketing research to model the purchase behavior of consumers.

9.4.2 Finite Mixtures of Logit and Probit Regression Models

Finite Mixture Regression Models for Binary Data

Let $Y_{i,t}$ denote a binary variable, observed for T_i times in reaction to a covariate $\mathbf{x}_i$, where the last element of $\mathbf{x}_i$ is 1 corresponding to an intercept. Define $Y_i = \sum_{t=1}^{T_i} Y_{i,t}$. It is assumed that the marginal distribution of Y_i follows a mixture of binomial distributions,

$$Y_i \sim \sum_{k=1}^{K} \eta_k \text{BiNom}\left(T_i, \pi_{k,i}\right), \tag{9.14}$$

where logit $\pi_{k,i} = \mathbf{x}_i\boldsymbol{\beta}_k$ in finite mixtures of logit regression models, whereas $\pi_{k,i} = \Phi(\mathbf{x}_i\boldsymbol{\beta}_k)$ in finite mixtures of probit regression models.

Both models capture extra-binomial variation due to unobserved heterogeneity in the population, for example, if an important covariate is omitted. It follows

$$\begin{aligned}
\text{E}(Y_i|\boldsymbol{\vartheta}) &= \pi_i = \sum_{k=1}^{K} \eta_k \pi_{k,i}, \\
\text{Var}(Y_i|\boldsymbol{\vartheta}) &= T_i\pi_i(1-\pi_i) + \frac{(T_i-1)T_i^2}{T_i}\left(\sum_{k=1}^{K} \eta_k \pi_{k,i}^2 - \pi_i^2\right).
\end{aligned} \tag{9.15}$$

For $T_i > 1$, extra-binomial variation due to the second term in (9.15) is present.

Identifiability is rather evolved for mixtures of logistic and probit regression models, the reason being that for $\mathbf{x}_i = 1$ such a mixture reduces to a

finite mixture of binomial distributions, which is not necessarily identifiable; see Subsection 9.3.1.

Pointwise identifiability for a fixed covariate $\mathbf{x}_i$ follows from the identifiability of a binomial mixture only, if $T_i \geq 2K - 1$. In this case

$$\sum_{k=1}^{K} \eta_k f_{BN}(y; T_i, \pi_{k,i}) = \sum_{k=1}^{K} \eta_k^\star f_{BN}(y; T_i, \pi_{k,i}^\star),$$

where logit $\pi_{k,i} = \mathbf{x}_i \boldsymbol{\beta}_k$ and logit $\pi_{k,i}^\star = \mathbf{x}_i \boldsymbol{\beta}_k^\star$, implies

$$\eta_k^\star = \eta_{\rho_i(k)}, \qquad \mathbf{x}_i \boldsymbol{\beta}_k^\star = \mathbf{x}_i \boldsymbol{\beta}_{\rho_i(k)}, \tag{9.16}$$

which is exactly what resulted for a Gaussian mixture of regression models. It follows immediately that a mixture of logistic regressions where only the intercept is switching is identifiable, if $T_i \geq 2K - 1$ for at least one covariate $\mathbf{x}_i$; see also Follmann and Lambert (1991).

Applications of finite mixtures of logistic regression models appear in biology to analyze the effect of a drug on the death rate of a protozoan trypanosome and to study the effect of salinity and temperature on the hatch rate of English sole eggs (Follmann and Lambert, 1989), in medicine to determine the risk factors of preterm delivery (Zhu and Zhang, 2004), in genetics to detect inheritance patterns of a binary trait such as alcoholism (Zhang and Merikangas, 2000), in marketing research to deal with the analysis of paired comparison choice data (Wedel and DeSarbo, 1993a), and in agriculture (Wang and Puterman, 1998).

Finite mixtures of probit regression models are applied in medical research to analyze the resistance to treatment of parasites in sheep (Lwin and Martin, 1989), in marketing research to analyze pick and/N data (De Soete and DeSarbo, 1991), and in the economics of labor markets (Geweke and Keane, 1999).

Finite Mixture Regression Models for Categorical Data

Extensions to multinomial mixtures are considered by Paul et al. (1989), Kim and Margolin (1992), and Morel and Nagaraj (1993). Kamakura and Russell (1989) applied a multinomial logit mixture regression model in marketing research to model consumers' choices among a set of brands and identified segments of consumers that differ in price sensitivity. Kamakura (1991) proposed a multinomial probit finite mixture regression model. Identifiability for multinomial mixture regression models is investigated in Grün (2002).

9.4.3 Parameter Estimation for Finite Mixtures of GLMs

ML estimation for finite mixtures of generalized linear models is considered, for instance, by Jansen (1993), Wedel and DeSarbo (1995), and Aitkin (1996).

Wedel et al. (1993), Lambert (1992), and Wang et al. (1996) use the EM algorithm for the estimation of mixtures of Poisson regression models, whereas Wedel and DeSarbo (1995) consider more general mixtures of GLMs.

For Bayesian estimation of mixtures of GLMs, Hurn et al. (2003) use the same prior as for a normal regression model, namely $\boldsymbol{\beta}_k \sim \mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$. Various suggestions have been put forward of how to estimate mixtures of nonnormal regression models using MCMC. Viallefont et al. (2002) use a single-move random walk Metropolis–Hastings algorithm, whereas Hurn et al. (2003) use a multivariate random walk Metropolis–Hastings algorithm for sampling directly from the marginal posterior distribution $p(\boldsymbol{\vartheta}|\mathbf{y})$ for mixtures of logistic and Poisson regressions; see also *Algorithm 3.6*. This is feasible, as the likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$ is available in closed form.

Alternatively, one could use data augmentation by introducing a group indicator S_i for each observation pair $(\mathbf{x}_i, \mathbf{y}_i)$ as missing data to obtain a sampling scheme comparable to *Algorithm 8.1*, which was derived in the context of finite mixtures of standard regression models. The resulting scheme, however, is not a Gibbs sampling scheme. Difficulties arise when drawing the group-specific parameters $\boldsymbol{\beta}_k$ in group k, because the conditional posterior distribution $p(\boldsymbol{\beta}_k|\mathbf{y}, \mathbf{S})$ has to be derived from a nonnormal regression model and does not belong to a well-known distribution family. To sample from this distribution, usually a Metropolis–Hastings step is applied; alternatively Hurn et al. (2003) mention the possibility of using the slice sampler (Damien et al., 1999). Classification, however, does not cause any problem, as it is sufficient to know the conditional distribution $p(\mathbf{y}_i|\boldsymbol{\beta}_k)$ for each $k = 1, \ldots, K$ for each observation $\mathbf{y}_i$.

9.4.4 Model Selection for Finite Mixtures of GLMs

Wang et al. (1996) use AIC and BIC for model selection of the number of mixture regressions, while including all possible covariates. Covariates are selected in a second step, after having chosen the number of components. In their simulation study BIC always selected the correct model.

Viallefont et al. (2002) use the reversible jump MCMC for mixtures of Poisson regression models to determine the number of mixture regressions, whereas Hurn et al. (2003) use the birth and death MCMC of Stephens (2000a). Both papers illustrate that the prior on K, which is usually assumed to be $\mathcal{P}(\lambda_0)$, is not without effects on the resulting inference.

9.5 Finite Mixture Models for Multivariate Binary and Categorical Data

In this section we consider finite mixture modeling of multivariate binary or categorical data $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir})$ is the realization of

an r-dimensional discrete random variable $\mathbf{Y} = (Y_1, \ldots, Y_r)$. Mixture models for multivariate discrete data, usually called latent class models, or latent structure analysis, have long been recognized as a useful tool in the behavioral and biomedical sciences, as exemplified by Lazarsfeld and Henry (1968), Goodman (1974b, 1978), Clogg and Goodman (1984), among many others; see Formann and Kohlmann (1996) and Clogg (1995) for a review.

In latent structure analysis the correlation between the elements $Y_1, \ldots, Y_r$ of $\mathbf{Y}$ is assumed to be caused by a discrete latent variable S_i, also called the latent class. It is then assumed that the variables $Y_1, \ldots, Y_r$, which are also called manifest variables, are stochastically independent conditional on the latent variable. Latent structure analysis is closely related to multivariate mixture modeling, as marginally the distribution of $\mathbf{Y}$ is a multivariate discrete mixture with density:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \prod_{j=1}^{r} p(y_{ij}|\boldsymbol{\pi}_{k,j}),$$

where $\boldsymbol{\pi}_{k,j}$ is a parameter modeling the discrete probability distribution of Y_j in class k.

9.5.1 The Basic Latent Class Model

In this section we consider a collection of multivariate binary observations $\mathbf{y}_1, \ldots, \mathbf{y}_N$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{ir})'$ is an r-dimensional vector of 0s and 1s, assumed to be the realization of a binary multivariate random variable $\mathbf{Y} = (Y_1, \ldots, Y_r)$. The latent class model assumes that associations between the manifest variables Y_j are caused by the presence of "latent classes" within which the features are independent. These latent classes may be seen as arising from an unobserved categorical variable S_i, which causes differences in occurrence probabilities $\pi_{k,j} = \Pr(Y_j = 1|S_i = k)$ of the manifest variable Y_j in the different classes k.

The marginal distribution of $\mathbf{Y}$ is equal to a mixture of r independent Bernoulli distributions, with density:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \prod_{j=1}^{r} \pi_{k,j}^{y_{ij}} (1 - \pi_{k,j})^{1-y_{ij}}, \tag{9.17}$$

where the K components of the mixture correspond to the K latent classes.

It is possible to verify that differences in the occurrence probabilities $\pi_{k,j}$ between the latent classes cause associations between the components of $\mathbf{Y}$ in the corresponding cell with respect to the marginal distribution, where the latent class is integrated out. For $K = 2$ and $r = 2$, for instance, Gilula (1979) shows that the marginal probability $\Pr(Y_1 = 1, Y_2 = 1|\boldsymbol{\vartheta})$ may be expressed as

$$\Pr(Y_1 = 1, Y_2 = 1|\boldsymbol{\vartheta}) =$$
$$\Pr(Y_1 = 1|\boldsymbol{\vartheta})\Pr(Y_2 = 1|\boldsymbol{\vartheta}) + \eta_1(1-\eta_1)(\pi_{1,1} - \pi_{2,1})(\pi_{1,2} - \pi_{2,2}).$$

Thus associations between the components of $\mathbf{Y}$ will be observed, whenever both occurrence probabilities are different. Bartholomew (1980) regards the latent class models as factor analysis for categorical data.

Historically seen, model (9.17) was originated by psychometricians and sociologists, and goes back to Lazarsfeld (1950). The main purpose was to study hypothetical constructs such as "intelligence" or "attitude." There is a large body of literature with many applications of these models to problems in behavioral, medical, and social sciences, such as finding associations between teaching style and pupil performance (Aitkin et al., 1981), tumor diagnostics based on a sequence of binary test results (Albert and Dodd, 2004), analyzing historical household data (Liao, 2004), and texture analysis (Grim and Haindl, 2003), just to mention a few.

Celeux and Govaert (1991) discuss the application of latent class models for clustering discrete data and, using the classification likelihood approach discussed earlier in Subsection 7.1.3, show that clustering based on the latent class model is closely related to clustering based on minimizing entropy-type criteria.

9.5.2 Identification and Parameter Estimation

A difficult problem with the latent class model is verifying if the model is identifiable for a given number of classes, given a certain collection of the data; see Goodman (1974b) and Clogg (1995). If $\pi_{k,1} = \cdots = \pi_{k,r} = \pi_k$, then a binomial finite mixture with component density $\text{BiNom}\,(r, \pi_k)$ results; consequently the more general latent class models could be applied only to at least three manifest variables ($r \geq 3$). As outlined by Formann and Kohlmann (1996, p.194), "In general it is not possible to say a priori whether these models may be identifiable or not." Statements about identifiability are usually made after having estimated the parameters under a certain model by considering the rank of the observed information matrix evaluated at the ML estimator as in Catchpole and Morgan (1997) to prove local identifiability (Rothenberg, 1971); see also Carreira-Perpiñán and Renals (2000).

In (9.17), the Kr unknown probabilities $(\pi_{1,1}, \ldots, \pi_{K,r})$ as well as the weight distribution $\boldsymbol{\eta}$ are unknown parameters that need to be estimated from the data. Pioneering work on ML estimation for the latent class model is found in Wolfe (1970). The basic latent class model is usually formulated as a generalized linear model and fitted by some iterative method, for instance, the proportional fitting algorithm of Goodman (1974a, 1974b), which later on turned out to be a variant of the EM algorithm.

An early reference on Bayesian estimation of latent class models is Evans et al. (1989), where the practical implementation was carried out using adaptive importance sampling. Again it is surprising to see how easily the marginal

posterior density $p(\pi_{1,1},\ldots,\pi_{K,r},\boldsymbol{\eta}|\mathbf{y})$, which is quite complicated, is obtained, using data augmentation and MCMC as in *Algorithm 3.4*. Assume that all probabilities $\pi_{k,j}$ are independent a priori, with $\pi_{k,j} \sim \mathcal{B}(a_{0,j}, b_{0,j})$. Conditional on the class indicator S_i, the conditional posterior $p(\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_K|\mathbf{S},\mathbf{y})$ is the product of Kr independent Beta distributions with

$$\pi_{k,j}|\mathbf{S},\mathbf{y} \sim \mathcal{B}\left(a_{0,j} + N_{k,j}(\mathbf{S}), b_{0,j} + N_k(\mathbf{S}) - N_{k,j}(\mathbf{S})\right),$$

where $N_{k,j}(\mathbf{S})$ is the number of ones observed for feature Y_j in latent class k, and $N_k(\mathbf{S})$ is the total number of observations in latent class k:

$$N_{k,j}(\mathbf{S}) = \sum_{i:S_i=k} y_{ij}, \qquad N_k(\mathbf{S}) = \#\{i : S_i = k\}.$$

9.5.3 Extensions of the Basic Latent Class Model

Over the years, many variants and extensions of the basic latent class model have been considered. One particularly useful extension deals with multivariate categorical data $\mathbf{y}_1,\ldots,\mathbf{y}_N$, where $\mathbf{y}_i = (y_{i1},\ldots,y_{ir})$ is the realization of an r-dimensional categorical random variable $\mathbf{Y} = (Y_1,\ldots,Y_r)$, with each element Y_j taking one value out of D_j categories $\{1,\ldots,D_j\}$. Again, a mixture density results:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k \prod_{j=1}^{r} \prod_{l=1}^{D_j} \pi_{k,jl}^{I_{\{y_{ij}=l\}}}, \tag{9.18}$$

where $\pi_{k,jl} = \Pr(Y_j = l|S_i = k)$ is the probability of category l for feature Y_j in class k.

The unknown parameter $\boldsymbol{\vartheta}$ appearing in (9.18) contains the unknown weight distribution $\boldsymbol{\eta}$ as well as the Kr unknown probability distributions $\boldsymbol{\pi}_{k,j} = (\pi_{k,j1},\cdots,\pi_{k,jD_j})$ of feature Y_j in class k. Again Bayesian estimation is easy implemented, by sampling from the marginal posterior density $p(\boldsymbol{\pi}_{1,1},\ldots,\boldsymbol{\pi}_{K,r}|\mathbf{y})$, using data augmentation and MCMC as in *Algorithm 3.4*. Assume that all probability distributions $\boldsymbol{\pi}_{k,j}$ are independent a priori, with $\boldsymbol{\pi}_{k,j} \sim \mathcal{D}(e_{0,j},\ldots,e_{0,j})$. Conditional on the class indicator S_i, the conditional posterior $p(\boldsymbol{\pi}_{1,1},\ldots,\boldsymbol{\pi}_{K,r}|\mathbf{S},\mathbf{y})$ is the product of Kr independent Dirichlet distributions with

$$\boldsymbol{\pi}_{k,j}|\mathbf{S},\mathbf{y} \sim \mathcal{D}\left(e_{0,j} + N_{k,j1}(\mathbf{S}),\ldots,e_{0,j} + N_{k,jD_j}(\mathbf{S})\right),$$

where, for each class k, $N_{k,jl}(\mathbf{S})$ counts how often category l is observed for feature j:

$$N_{k,jl}(\mathbf{S}) = \sum_{i:S_i=k} I_{\{y_{ij}=l\}}.$$

In a series of papers, Formann (1982, 1992, 1993, 1994a, 1994b) considered further extensions of (9.18) such as linear logistic latent class analysis, where both η_k as well as the probabilities $\pi_{k,jl}$ depend on some covariates through a linear logistic model; see also Formann and Kohlmann (1996) for a review.

Clogg and Goodman (1984) consider simultaneously a latent structure analysis of a whole set of multinomial contingency tables and discuss methods for testing complete or partial homogeneity across tables.

9.6 Further Issues

9.6.1 Finite Mixture Modeling of Mixed-Mode Data

Often data are realizations of a mixed random variable $\mathbf{Y} = (\mathbf{Y}^C, \mathbf{Y}^D)$ with $\mathbf{Y}^C$ containing metric features and $\mathbf{Y}^D$ containing categorical features. Within a latent class analysis of such mixed-mode data, it is assumed that the distribution of $\mathbf{Y}$ depends on a latent unknown variable, which again leads to a finite mixture model.

Everitt (1988) and Everitt and Merette (1990) deal with mixed-mode data by incorporating the use of thresholds for the categorical data, however, the resulting model is difficult to estimate. Muthén and Shedden (1999) suggest combining features of Gaussian multivariate mixtures with a latent class model. The density of this mixture model reads:

$$p(\mathbf{y}_i|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k f_N(\mathbf{y}_i^C; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) p(\mathbf{y}_i^D|\boldsymbol{\theta}_k^D), \tag{9.19}$$

$$p(\mathbf{y}_i^D|\boldsymbol{\theta}_k^D) = \prod_{j=1}^{r} \prod_{l=1}^{D_j} \pi_{k,jl}^{I_{\{y_{ij}^D = l\}}}.$$

Bacher (2000) discusses an application of this model in sociology. Clustering multivariate data through probabilistic models based on finite mixtures is particularly useful for mixed continuous and categorical data; see Bock (1996) for some review.

Muthén and Shedden (1999) use the EM algorithm for estimation. Bayesian estimation of model (9.19) using data augmentation and MCMC as in *Algorithm 3.4* is easily implemented, as conditionally on a known classification we only need to combine the sampling step for multivariate mixtures of normals discussed in Subsection 6.3.3 with those obtained for the latent class model in Subsection 9.5.3.

A disadvantage of model (9.19) is stochastic independence of the categorical and the continuous variables within each class; more refined models, that allow for association between both types of variables are discussed in Lawrence and Krzanowski (1996). The idea is to replace the multivariate categorical variable $\mathbf{Y}^D = (Y_1^D, \ldots, Y_r^D)$, each of which is assumed to have D_j different

categories by a single multinomial variable $\mathbf{Y}^M$, which has $D_M = \prod_{j=1}^{r} D_j$ different cells, corresponding to the number of distinct patterns produced by $\mathbf{Y}^D$. Furthermore the group-specific mean $\boldsymbol{\mu}_{k,l}$ of the continuous variable is allowed to be different for all patterns $l = 1, \ldots, D_M$.

Willse and Boik (1999) show that the model in its unrestricted form is not identifiable. There exist $(K!)^{D_M - 1}$ distinct parameters that define the same mixture distribution. Identifiability is achieved by imposing the restrictions $\boldsymbol{\mu}_{k,l} = \boldsymbol{\mu}_k + \boldsymbol{\beta}_l$ on the group-specific mean of the continuous variable.

Hunt and Jorgensen (2003) extended model (9.19) to mixed-mode data with missing observations. The modeling of mixed-mode data in a time series context is discussed in Cosslett and Lee (1985).

9.6.2 Finite Mixtures of GLMs with Random Effects

As discussed in Section 9.4, finite mixtures of GLMs are able to deal with overdispersion and extra-binomial or multinomial variation in regression models for discrete valued data. An alternative popular approach is based on GLMs with random-effects models (Schall, 1991; Breslow and Clayton, 1993; Aitkin, 1996), which regard overdispersion and extra-binomial or multinomial variation as a nuisance factor that needs to be accounted for in order to obtain consistent estimates of the other parameters. GLMs with random effects are also applied to pool information across similar units as in Section 8.5 for repeated measurements where the dependent variable is a discrete rather than a normally distributed random variable.

Usually the distribution of the random effects is chosen to be normal. Neuhaus et al. (1992) studied the effect of misspecifying the distribution of the random effects for logistic mixed-effects models and found cases of inconsistency both for the fixed and the random effects. Much more flexibility is achieved by assuming that the random effects follow a mixture of normal distributions as in Section 8.5, in which case a finite mixture of GLMs with random effects results. Such a model has been applied by Lenk and DeSarbo (2000) in marketing research, who discuss Bayesian estimation using data augmentation and MCMC. Yau et al. (2003) apply a two-component mixture of binary logit models with random effects to the analysis of hospital length of stay. Bottolo et al. (2003) apply a mixture of Poisson models with random effects to modeling extreme values in a data set of large insurance claims, using reversible jump MCMC.