23
Non-Gaussian Random Effects

23.1 Introduction

The mixed models discussed so far all assume that the random effects are normally distributed. This assumption has been carried over from the linear mixed models, where it has proven to be mathematically very convenient in the sense that the marginal likelihood can easily be calculated analytically (Chapter 4). In non-linear mixed models, as well as in generalized linear mixed models, this normality assumption has been the cause of many computational difficulties because the marginal likelihood can no longer be computed analytically, which has resulted in many proposals in the statistical literature about how to approximate the likelihood to be maximized (see Chapter 14 for an overview).

For linear mixed models, it has been shown (Verbeke and Lesaffre 1996, 1997) that deviations from this normality assumption have very little impact on the estimation of the parameters in the marginal model, but much more on the empirical Bayes estimates for the random effects. For non-linear and generalized linear mixed models, misspecification of the random-effects distribution can lead to biased estimates for the parameters in the marginal model, including the fixed effects that are usually of primary interest. We refer to Neuhaus, Hauck, and Kalbfleisch (1992), Butler and Louis (1992), Pfeiffer *et al* (2003), Heagerty and Zeger (2000), and Litiére *et al* (2005) for more details on the effect of misspecifications of random-effects distributions in generalized linear mixed models.

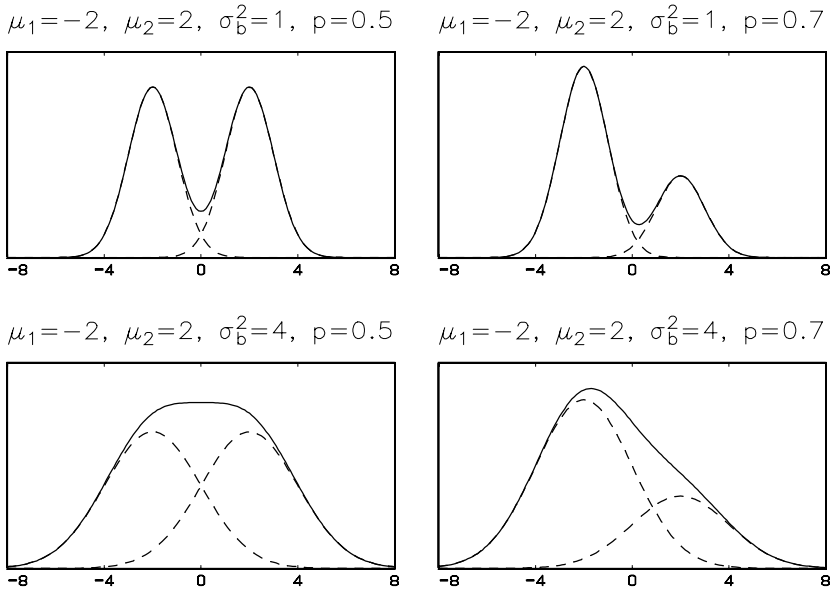

FIGURE 23.1. *Density functions of mixtures* $pN(\mu_1, \sigma_b^2)+(1-p)N(\mu_2, \sigma_b^2)$ *of two normal distributions, for varying values for* p *and* σ_b^2*. The dashed lines represents the densities of the normal components; the solid line represents the density of the mixture.*

This calls for methods to check the normality of the random effects and for models that relax the distributional assumptions. In the context of linear mixed models, it has been shown (Verbeke and Molenberghs 2000 Section 7.8) that the empirical Bayes estimates for the random effects, obtained under normality, cannot be used to check normality because the prior belief of normality often forces the estimates to satisfy this assumption such that non-normality of the random effects may not be reflected in their empirical Bayes estimates. Therefore, Verbeke and Lesaffre (1996), Magder and Zeger (1996), and Verbeke and Molenberghs (2000, Chapter 12) have extended the linear mixed model with mixtures of normals as random-effects distribution. This particular extension has several advantages. First, as shown in Figure 23.1, the class of finite mixtures of normal distributions is a very flexible class of distributions: unimodal as well as multimodal, symmetric as well as very skewed. Second, mixtures can be used to model unobserved heterogeneity in the random-effects distribution. Third, the fact that the mixture components are still normally distributed allows the implementation to take advantage of algorithms and software already available for fitting the models with normally distributed random effects. Finally, the mixture models can be used for classification purposes, which makes them

particularly useful in contexts of discriminant analysis or cluster analysis, based on longitudinal profiles.

In this chapter, we will present and illustrate the mixture approach in the context of generalized linear, or non-linear, mixed models. In Section 23.2, the model will be introduced. In Section 23.3, estimation and inference will be discussed. Section 23.4 briefly explains how random effects can be estimated under the mixture assumption and shows how the mixture models can be used for classification purposes. Finally, an example will be worked out in Section 23.5. More details on the model, as well as on the related estimation and inference can be found in Fieuws, Spiessens, and Draney (2004) or in Muthén and Shedden (1999).

23.2 The Heterogeneity Model

As before, let $\boldsymbol{Y}_i$ be the n_i-dimensional vector of all measurements available for cluster $i = 1, \ldots, N$, and let $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$ be the corresponding density, conditional on a q-dimensional vector $\boldsymbol{b}_i$ of random effects. We hereby do not explicitly denote possible dependence of $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$ on unknown parameters such as fixed effects. In the mixed models considered so far, the random effects $\boldsymbol{b}_i$ were always assumed to be sampled from a normal distribution with mean vector zero and a covariance matrix D, i.e., $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$. This assumption reflects the prior believe that the random effects are drawn from one homogeneous population of random effects. From now on, the so-obtained mixed model will be termed 'homogeneity' model.

The 'heterogeneity' model is obtained by replacing the normality assumption for the random effects by a mixture of g q-dimensional normal distributions with mean vectors $\boldsymbol{\mu_r}$ and covariance matrices D_r, i.e.,

$$\boldsymbol{b_i} \;\sim\; \sum_{r=1}^{g} p_r N(\boldsymbol{\mu_r}, D_r), \tag{23.1}$$

with $\sum_{r=1}^{g} p_r = 1$. The population under study can then be interpreted as a combination of g sub-populations, each representing a fraction p_r of the total population. In the rth sub-population, the random effects are normally distributed with mean $\boldsymbol{\mu_r}$, and covariance D_r. Clearly, model (23.1) reflects prior belief of presence of unobserved heterogeneity. Therefore, the resulting mixed model is called 'heterogeneity' model.

We now define $z_{ir} = 1$ if $\boldsymbol{b}_i$ is sampled from the rth component in the mixture, and 0 otherwise, $r = 1, \ldots, g$. We then have that $P(z_{ir} = 1) = E(z_{ir}) = p_r$ and that

$$E(\boldsymbol{b}_i) = E\left[E(\boldsymbol{b}_i \mid z_{i1}, \ldots, z_{ig})\right] = E\left(\sum_{r=1}^{g} \boldsymbol{\mu}_r \; z_{ir}\right) = \sum_{r=1}^{g} p_r \; \boldsymbol{\mu}_r.$$

Therefore, the additional constraint $\sum_{r=1}^{g} p_r \boldsymbol{\mu}_r = 0$ is needed to ensure that the random effects still have mean zero. Further, we have that the overall covariance matrix of the $\boldsymbol{b}_i$ is given by

$$\begin{aligned} D^* &= \text{var}\left[E(\boldsymbol{b}_i \mid z_{i1}, \ldots, z_{ig})\right] + E\left[\text{var}(\boldsymbol{b}_i \mid z_{i1}, \ldots, z_{ig})\right] \\ &= \text{var}\left(\sum_{r=1}^{g} \boldsymbol{\mu}_r \; z_{ir}\right) + E\left(\sum_{r=1}^{g} D_r \; z_{ir}\right) \\ &= \sum_{r=1}^{g} p_r \boldsymbol{\mu}_r \boldsymbol{\mu}_r' + \sum_{r=1}^{g} p_r D_r. \end{aligned} \tag{23.2}$$

The first term represents variability between the mixture components, and the second term is the average within-component variability. Hence, (23.2) can be interpreted as a decomposition of variability in the random effects in terms of variability between and variability within the sub-populations. Finally, denoting the density within the rth mixture component by $f_r(\boldsymbol{b}_i)$, we have that the density function corresponding to (23.1) is given by

$$\begin{aligned} f(\boldsymbol{b}_i) &= \sum_{r=1}^{g} p_r f_r(\boldsymbol{b}_i) \\ &= \sum_{r=1}^{g} p_r \; (2\pi)^{-q/2} \; |D_r|^{-1/2} \\ &\qquad \times \; \exp\left\{-\frac{1}{2}\left(\boldsymbol{b}_i - \boldsymbol{\mu}_r\right)' D_r^{-1} \left(\boldsymbol{b}_i - \boldsymbol{\mu}_r\right)\right\}. \end{aligned} \tag{23.3}$$

It should be emphasized that we consider the number of components g in (23.1) to be known. In practice, several models can be fitted, with increasing values for g, leading to a series of nested models, and testing procedures such as the likelihood ratio test could be used for the comparison of these models. However, as discussed by Ghosh and Sen (1985), testing for the number of components in a finite mixture is seriously complicated by boundary problems similar to the ones discussed in Section 14.6 in the context of tests for variance components. In order to briefly highlight the main problems, we consider testing $H_0 : g = 1$ versus $H_A : g = 2$. The null hypothesis can then be expressed as $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. However, the same hypothesis is obtained by setting $H_0 : p_1 = 0$ or $H_0 : p_2 = 0$, which clearly illustrates that H_0 is on the boundary of the parameter space, and hence also that the usual regularity conditions for application of the classical maximum likelihood theory are violated. Therefore, simulations are needed to derive the correct null distribution of the LR test statistic. We refer to Verbeke (1995, Section 4.6) for an example in the context of linear mixed models, and to McLachlan and Basford (1988, Section 1.10) for an extensive overview of the literature on the use of the LR test in finite

mixture problems. In practice it is often sufficient to fit several heterogeneity models and to explore how increasing g affects the inference for the parameters of interest.

In the context of linear mixed models, Magder and Zeger (1996) also considered mixtures of normal distributions as random-effects distribution, but they treated the number g of components as an unknown parameter, to be estimated from the data. In order to avoid that non-smooth mixture distributions, with many components, would be obtained, they pre-specify a lower boundary h for the within-component variability measured by the determinants $|D_r|$ of the within-component covariance matrices. In practice, very little difference is expected from models that pre-specify the number of mixture components. Indeed, when a very smooth mixing distribution is required, a large value of h can be specified, which will yield a mixture of a relatively small number of normal distributions.

23.3 Estimation and Inference

Estimation and inference for the heterogeneity model will be based on maximum likelihood (ML) principles for the marginal likelihood of the data. The marginal distribution of $\boldsymbol{Y}_i$, obtained from integrating out the random effects, is given by

$$\begin{aligned} f_i(\boldsymbol{y}_i) &= \int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)\ f(\boldsymbol{b}_i)\ d\boldsymbol{b}_i \\ &= \sum_{r=1}^{g} p_r \int f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)\ f_r(\boldsymbol{b}_i)\ d\boldsymbol{b}_i \\ &= \sum_{r=1}^{g} p_r f_{ir}(\boldsymbol{y}_i) \end{aligned} \tag{23.4}$$

in which $f_{ir}(\boldsymbol{y}_i)$ is the marginal density corresponding to a mixed model with random effects that are normally distributed with mean $\boldsymbol{\mu_r}$ and covariance D_r. Hence, the marginal density of $\boldsymbol{Y}_i$ is again a g-component finite mixture, with the same mixing proportions p_r, and where the component-specific densities are marginal mixed model densities within the specific sub-population. This specific feature will simplify implementation considerably because it will be possible to build on existing software for generalized linear and/or non-linear mixed models.

Maximization of the marginal likelihood resulting from (23.4) will be based on the so-called Expectation-Maximization (EM) algorithm, see Laird (1978). See also Section 28.3 for a general introduction of the algorithm in the context of missing data. The EM algorithm is particularly useful for mixture problems because it often happens that a model is fitted with too many components (g too large), leading to a likelihood that is maximal

anywhere on a ridge. As shown by Dempster, Laird, and Rubin (1977), the EM algorithm is capable of converging to some particular point on that ridge. Titterington, Smith, and Makov (1985, pp. 88–89) compare the EM algorithm with the Newton-Raphson (NR) algorithm. Their conclusions can be summarized as follows:

- EM is usually simple to apply and satisfies the appealing monotonic property in that it increases the objective function at each iteration step. NR is more complicated, and there is no guarantee of monotonicity.
- If NR converges, it is of second order (i.e., fast), whereas EM is often painfully slow. However, if the separation between the components in the mixture is poor, even the numerical performance of NR can be disappointing. Simulations have shown that, in such cases, NR can fail to converge in up to half the simulations, even when the algorithm was started from the true parameter values.
- Convergence is not guaranteed with any of the techniques because EM, even with the monotonicity property, can converge to a local maximum of the likelihood surface.

Böhning and Lindsay (1988) have considered maximization of log-likelihoods for which the quadratic approximation based on the Taylor series is "flatter" than the objective function, thereby sending the solution too far at the next step. They conclude that, in a mixture framework, flat log-likelihoods often occur. It is known that this often leads to problems in convergence and to instabilities for the Newton-Raphson algorithm.

Note also that because the random effects are assumed to follow a mixture of distributions of the same parametric family, the vector of all parameters in the marginal model is, strictly speaking, not identifiable. Indeed, the log-likelihood is invariant under the $g!$ possible permutations of the mean vectors $\boldsymbol{\mu_r}$, the covariances D_r, and the corresponding component probabilities p_r. Therefore, the likelihood will have at least $g!$ local maxima with the same likelihood value. However, this lack of identifiability is of no concern in practice, as it can easily be overcome by imposing some constraint on the parameters. For example, Aitkin and Rubin (1985) use the constraint that

$$p_1 \geq p_2 \geq \ldots \geq p_g. \tag{23.5}$$

The likelihood is then maximized without the restriction, and the component labels are permuted afterwards to achieve (23.5).

The EM algorithm is frequently used for the calculation of maximum likelihood estimates for missing data problems (Section 28.3). Strictly speaking, we do not necessarily have missingness in our context. However, it will prove extremely convenient to treat the component membership indicators

z_{ir}, $i = 1, \ldots, N$, $r = 1, \ldots, g$ as missing. We now give a brief introduction on the EM algorithm in the context of the heterogeneity model, and we refer to McLachlan and Basford (1988, Section 1.6) for an application of the EM algorithm in a simpler mixture context, where it is assumed that the available data are all drawn from the same mixture distribution (no different dimensions, no covariates).

Let $\boldsymbol{\pi}$ be the vector of component probabilities [i.e., $\boldsymbol{\pi}' = (p_1, \ldots, p_g)$] and let $\boldsymbol{\gamma}$ be the vector containing the remaining parameters, i.e., the parameters in the conditional densities $f_i(\boldsymbol{y}_i|\boldsymbol{b}_i)$ as well as in all $\boldsymbol{\mu_r}$ and all D_r. Further, $\boldsymbol{\theta}' = (\boldsymbol{\pi}', \boldsymbol{\gamma}')$ denotes the vector of all parameters in the marginal heterogeneity model (23.4). Further, we now explicitly denote dependence of the within-component marginal densities $f_{ir}(\boldsymbol{y}_i)$ on $\boldsymbol{\gamma}$ by $f_{ir}(\boldsymbol{y}_i|\boldsymbol{\gamma})$. The marginal likelihood function is then given by

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{N} \left[\sum_{r=1}^{g} p_r \; f_{ir}(\boldsymbol{y}_i \mid \boldsymbol{\gamma}) \right], \tag{23.6}$$

where $\boldsymbol{y}' = (\boldsymbol{y_1}', \ldots, \boldsymbol{y_N}')$ is the vector containing all observed response values.

Let z_{ir} be as defined before in Section 23.2. The prior probability for an individual to belong to component r is then $P(z_{ir} = 1) = p_r$, the mixture proportion for that component. The log-likelihood function for the observed measurements $\boldsymbol{y}$ and for the vector $\boldsymbol{z}$ of all unobserved z_{ir} is then

$$\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{N} \sum_{r=1}^{g} z_{ir} \left[\ln p_r \; + \; \ln f_{ir}(\boldsymbol{y_i}|\boldsymbol{\gamma}) \right],$$

which is easier to maximize than the log-likelihood function corresponding to the likelihood (23.6) of the observed data vector $\boldsymbol{y}$ only. On the other hand, maximizing $\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$ with respect to $\boldsymbol{\theta}$ yields estimates which depend on the unobserved ("missing") indicators z_{ir}. A compromise is obtained with the EM algorithm, where the expected value of $\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$, rather than $\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$ itself, is maximized with respect to $\boldsymbol{\theta}$, where the expectation is taken over all the unobserved z_{ir}. In the E step (expectation step), the conditional expectation of $\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z})$, given the observed data vector $\boldsymbol{y}$, is calculated. In the M step (maximization step), the so-obtained expected log-likelihood function is maximized with respect to $\boldsymbol{\theta}$, providing an updated estimate for $\boldsymbol{\theta}$. Finally, one keeps iterating between the E step and the M step until convergence is attained.

More specifically, let $\boldsymbol{\theta}^{(t)}$ be the current estimate for $\boldsymbol{\theta}$, and $\boldsymbol{\theta}^{(t+1)}$ stands for the updated estimate, obtained from one further iteration in the EM algorithm. We then have the following E and M steps in the estimation process for the heterogeneity model.

The E Step. The conditional expectation

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \quad = \quad E\left[\ell(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}) \mid \boldsymbol{y}, \boldsymbol{\theta}^{(t)} \right]$$

is given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \;=\; \sum_{i=1}^{N}\sum_{r=1}^{g} p_{ir}(\boldsymbol{\theta}^{(t)})\left[\ln p_r \;+\; \ln f_{ir}(\boldsymbol{y_i}|\boldsymbol{\gamma})\right], \quad (23.7)$$

where only the posterior probability for the ith individual to belong to the rth component of the mixture, given by

$$\begin{aligned} p_{ir}(\boldsymbol{\theta}^{(t)}) &= E(z_{ir} \mid \boldsymbol{y_i}, \boldsymbol{\theta}^{(t)}) = P(z_{ir} = 1 \mid \boldsymbol{y}_i, \boldsymbol{\theta}^{(t)}) \\ &= \left.\frac{p_r f_{ir}(\boldsymbol{y_i}|\boldsymbol{\gamma})}{\sum_{k=1}^{g} p_k f_{ik}(\boldsymbol{y_i}|\boldsymbol{\gamma})}\right|_{\hat{\pi}^{(t)},\hat{\gamma}^{(t)}} \end{aligned}$$

has to be calculated for each i and r.

The M Step. To get the updated estimate $\boldsymbol{\theta}^{(t+1)}$, we have to maximize expression (23.7) with respect to $\boldsymbol{\theta}$. We first maximize

$$\begin{aligned} &\sum_{i=1}^{N}\sum_{r=1}^{g} p_{ir}(\boldsymbol{\theta}^{(t)}) \ln p_r \\ &\quad = \sum_{i=1}^{N}\sum_{r=1}^{g-1} p_{ir}(\boldsymbol{\theta}^{(t)}) \ln p_r \;+\; \sum_{i=1}^{N} p_{ig}(\boldsymbol{\theta}^{(t)}) \;\ln\left(1 - \sum_{r=1}^{g-1} p_r\right) \end{aligned}$$

with respect to $p_1, \ldots, p_{g-1}$. Setting all first-order derivatives equal to zero establishes that the updated estimates satisfy

$$\frac{p_r^{(t+1)}}{p_g^{(t+1)}} \;=\; \frac{\sum_{i=1}^{N} p_{ir}(\boldsymbol{\theta}^{(t)})}{\sum_{i=1}^{N} p_{ig}(\boldsymbol{\theta}^{(t)})},$$

for all $r = 1, \ldots, g-1$. This also implies that

$$1 \;=\; \sum_{r=1}^{g} p_r^{(t+1)} \;=\; \frac{N\; p_g^{(t+1)}}{\sum_{i=1}^{N} p_{ig}(\boldsymbol{\theta}^{(t)})},$$

from which it follows that all estimates $p_r^{(t+1)}$ satisfy

$$p_r^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} p_{ir}(\boldsymbol{\theta}^{(t)}).$$

Unfortunately, the second part of (23.7) cannot be maximized analytically, and a numerical maximization procedure such as Newton-Raphson is needed to maximize

$$\sum_{i=1}^{N}\sum_{r=1}^{g} p_{ir}(\boldsymbol{\theta}^{(t)}) \ln f_{ir}(\boldsymbol{y_i}|\boldsymbol{\gamma}) \quad (23.8)$$

with respect to $\boldsymbol{\gamma}$. Luckily, (23.8) can be interpreted as a weighted log-likelihood of a generalized linear or non-linear mixed model. Therefore, maximization of (23.8), can often be based on software procedures available for fitting generalized linear and non-linear mixed models, such as the SAS procedures GLIMMIX and NLMIXED (Chapter 15). We refer to Fieuws, Spiessens, and Draney (2004) for an implementation based on the NLMIXED procedure.

Often, numerical maximization algorithms are based on second-order derivatives of the log-likelihood function. This allows easy calculation of the observed Fisher information matrix and hence also of asymptotic standard errors for all ML estimates. This is not the case for the EM algorithm, which immediately highlights one of the main drawbacks of this algorithm. However, Louis (1982) has provided a procedure for approximating the observed information matrix with few additional calculations. The so-obtained standard errors can then be used to construct classical asymptotic Wald-type tests, based on the asymptotic normality of the ML estimators. Alternative inferences can be based on likelihood ratio principles as well.

23.4 Empirical Bayes Estimation and Classification

When the random effects $\boldsymbol{b}_i$ are of interest, empirical Bayes (EB) techniques can be used for their estimation. As explained in Section 14.2.4, it is customary to define the EB estimates as the posterior modes of the random effects $\boldsymbol{b}_i$, i.e., as the value for $\boldsymbol{b}_i$ that maximizes the posterior density $f_i(\boldsymbol{b_i}|\boldsymbol{y}_i)$, in which all unknown parameters have been replaced by their estimates obtained from maximizing the marginal likelihood function. Under the heterogeneity model, the posterior density of $\boldsymbol{b}_i$ is given by

$$f_i(\boldsymbol{b}_i \mid \boldsymbol{y_i}, \boldsymbol{\theta}) \quad = \quad \sum_{r=1}^{g} p_{ir}(\boldsymbol{\theta}) f_{ir}(\boldsymbol{b_i} \mid \boldsymbol{y_i}, \boldsymbol{\gamma}), \tag{23.9}$$

where $f_{ir}(\boldsymbol{b}_i|\boldsymbol{y}_i, \boldsymbol{\gamma})$ is the posterior density function of $\boldsymbol{b}_i$, conditional on $z_{ir} = 1$, i.e., conditional on the knowledge that $\boldsymbol{b}_i$ was sampled from the rth component in the mixture. Hence, the posterior distribution of $\boldsymbol{b_i}$ is a mixture of the posterior distributions of $\boldsymbol{b}_i$ within each component of the mixture, with the posterior probabilities $p_{ir}(\boldsymbol{\theta})$ as subject-specific mixture proportions. The possible multimodality of the posterior density of $\boldsymbol{b}_i$ implies that the posterior mode is not a good point estimate for $\boldsymbol{b}_i$, in many applications. However, expression (23.9) suggests estimating the random effect $\boldsymbol{b}_i$ for cluster i by the weighted sum

$$\widehat{\boldsymbol{b}_i} \quad = \quad \sum_{r=1}^{g} p_{ir}(\boldsymbol{\theta}) \widehat{\boldsymbol{b_{ir}}}(\boldsymbol{\gamma})$$

of the component-specific posterior modes $\widehat{\boldsymbol{b_{ir}}}(\boldsymbol{\gamma})$, with weights equal to the posterior probabilities for that subject to belong to the different mixture components, and with the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ replaced by their ML estimates obtained from the EM algorithm. The resulting estimates will still be called empirical Bayes estimates.

Interest could also lie in the classification of the subjects into the different mixture components. It is natural in mixture models for such a classification to be based on the estimated posterior probabilities $p_{ir}(\widehat{\boldsymbol{\theta}})$ (McLachlan and Basford 1988, Section 1.4). One then classifies the ith subject into the component for which it has the highest estimated posterior probability to belong to, that is, to the $r(i)$th component, where $r(i)$ is the index for which

$$p_{i,r(i)}(\widehat{\boldsymbol{\theta}}) = \max_{1 \leq r \leq g} p_{ir}(\widehat{\boldsymbol{\theta}}).$$

Note how this technique can be used for cluster analysis within the framework of non-linear or generalized linear mixed models: If the individual profiles are to be classified into g subgroups, fit a mixture model with g components and use the above rule for classification in either one of the g clusters. In the context of discriminant analysis, a mixed model can be fitted to each group separately, and a mixture model can be used for the classification of future clusters. Examples in the context of linear models for continuous data can be found in Verbeke and Lesaffre (1996), Tomasko, Helms, and Snapinn (1999), Verbeke and Molenberghs (2000, Chapter 12), and Brant *et al* (2003). An example in the context of non-linear mixed models can be found in Fieuws, Verbeke, and Brant (2005).

23.5 The Verbal Aggression Data

As an illustration of the mixture approach, we re-analyze the data of Vansteelandt (2000), which were also used by De Boeck and Wilson (2004), as key example throughout their whole book. The data are responses from 316 persons to questions (items) about verbal aggression. All items refer to verbally aggressive reactions in a frustrating situation. For example, one item is: 'A bus fails to stop for me. I would curse.' Possible responses are 'Yes,' or 'No.' Further, the experimental design has four factors, summarized in Table 23.1. The first one is the type of behavior, with possible values 'Curse,' 'Scold,' and 'Shout.' The second design factor is the behavior mode. A differentiation is made between actual doing (i.e., cursing, scolding, or shouting) and wanting to do (i.e., wanting to curse, wanting to scold, or wanting to shout). The third design factor is the situation type. This factor has two levels: situations in which someone else is to blame, and situations in which one is self to blame. Examples of other-to-blame situations are 'A bus fails to stop for me,' and 'I miss a train because a clerk gave me faulty information.' Examples of self-to-blame situations are 'The

TABLE 23.1. *Verbal Aggression Data. Summary of the 24 items. Two versions exist of each item. The version with 'want to' in the item formulation refers to items with behavior mode 'Want.' The version without 'want to' in the item formulation refers to items with behavior mode 'Do.'*

Items		Situation type	Behavior
1.	A bus fails to stop for me. I would (want to) curse.	Other to blame	Curse
2.	A bus fails to stop for me. I would (want to) scold.		Scold
3.	A bus fails to stop for me. I would (want to) shout.		Shout
4.	I miss a train because a clerk gave me faulty information. I would (want to) curse.		Curse
5.	I miss a train because a clerk gave me faulty information. I would (want to) scold.		Scold
6.	I miss a train because a clerk gave me faulty information. I would (want to) shout.		Shout
7.	The grocery store closes just as I am about to enter. I would (want to) curse.	Self to blame	Curse
8.	The grocery store closes just as I am about to enter. I would (want to) scold		Scold
9.	The grocery store closes just as I am about to enter. I would (want to) shout.		Shout
10.	The operator disconnects me when I had used up my last 10 cents for a call. I would (want to) curse.		Curse
11.	The operator disconnects me when I had used up my last 10 cents for a call. I would (want to) scold.		Scold
12.	The operator disconnects me when I had used up my last 10 cents for a call. I would (want to) shout.		Shout

operator disconnects me when I had used up my last 10 cents for a call,' and 'The grocery store closes just as I am about to enter.' The fourth factor, the specific situations that are asked about (2 of each-see Table 23.1), is nested within the third. This factor will not be used in the analyses here. In conclusion, the design is a $3 \times 2 \times 2$ design with a fourth factor nested within the third, with 24 items in total.

Let Y_{ij} be the outcome for the jth item, measured on respondent i, $i = 1, \ldots, 316$, $j = 1, \ldots, 24$. Further, we define four dummy variables, as defined in Table 23.2. The definition of X_2 and X_3 is such that they characterize expression of frustration (X_2) and expression of blame (X_3). In our analyses, we will focuss on the effect of the factor 'Type of situation,' and more specifically, to the heterogeneity in the population with respect to the effect this factor has on the outcome. All our models will be of the

TABLE 23.2. *Verbal Aggression Data. Definition of the dummy variables for the design factors.*

Dummy	Design factor	Definition
X_1:	Type of situation:	$\begin{cases} X_1 = 1 & \text{Other to blame} \\ X_1 = 0 & \text{Self to blame} \end{cases}$
X_2, X_3:	Type of behavior:	$\begin{cases} X_2 = 1 & \text{Cursing or shouting} \\ X_2 = 0 & \text{Scolding} \end{cases}$
		$\begin{cases} X_3 = 1 & \text{Cursing or scolding} \\ X_3 = 0 & \text{Shouting} \end{cases}$
X_4:	Mode of behavior:	$\begin{cases} X_4 = 1 & \text{Do mode} \\ X_4 = 0 & \text{Want mode} \end{cases}$

form

$$\begin{aligned} Y_{ij}|\boldsymbol{b}_i &\sim \text{Bernoulli}(\pi_{ir}), \\ \text{logit}(\pi_{ir}) &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})X_{1i} \\ &\quad +\beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}, \end{aligned} \tag{23.10}$$

in which $\boldsymbol{b}_i = (b_{i0}, b_{i1})'$ represents the vector of random (subject-specific) intercepts and random (subject-specific) effects of 'Others to blame' (X_1). It is assumed that the random effects $\boldsymbol{b}_i$ satisfy

$$\boldsymbol{b}_i \sim \sum_{r=1}^{g} p_r N(\boldsymbol{\mu_r}, D_r),$$

where, as before $\sum_r p_r \boldsymbol{\mu_r} = \mathbf{0}$. Here, we will only consider models with the same covariance matrix in all mixture components, i.e., with all D_r equal to D,

$$\boldsymbol{b}_i \sim \sum_{r=1}^{g} p_r N\left[\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix}, \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}\right],$$

where $\boldsymbol{\mu_r} = (\mu_{0j}, \mu_{1j})'$.

Depending on the actual form of the $\boldsymbol{\mu_r}$ and of D, we get a variety of models all known in the psychometric literature. We refer to Fieuws, Spiessens, and Draney (2004) for a detailed discussion. A graphical representation of several of those models is given in Figure 23.2, in case of two mixture components, i.e., $g = 2$. For example, if the within-component covariance D is the 2×2 zero matrix, then no within-component variability is present, and Model (23.10) reduces to a so-called latent class model,

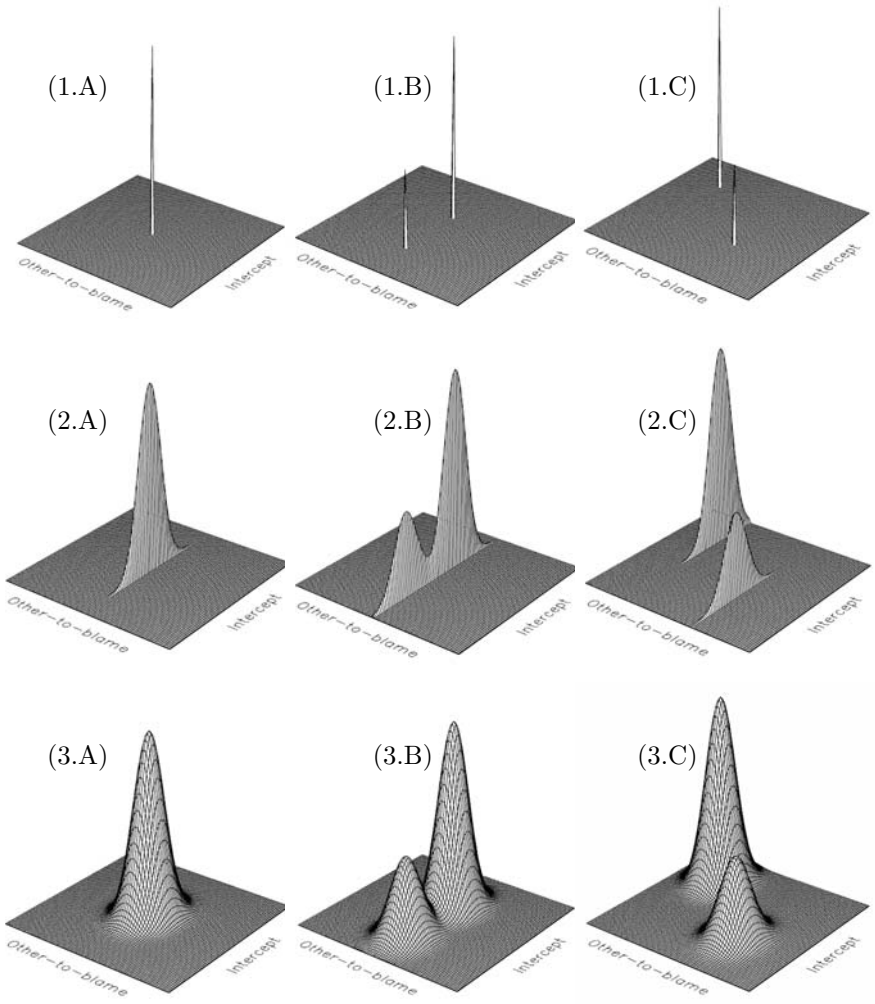

FIGURE 23.2. *Verbal Aggression Data. Graphical representation of different distributional assumptions for random effects.*
Classification according to amount of variability within the mixture components:
Row 1: no variability
Row 2: only variability for intercepts
Row 3: variability for intercepts and effects of other to blame
Classification according to discrimination of the mixture components:
Column A: no discrimination at all
Column B: discrimination on intercepts only
Column C: discrimination on intercepts and effects of other to blame

TABLE 23.3. *Verbal Aggression Data. Maximum likelihood estimates (standard errors) for a one-component and several two-component mixture models.*

		Heterogeneity models ($g = 2$)		
Effect	Homogeneity	Model A	Model B	Model C
β_0	-0.31 (0.096)			-0.32 (0.06)
$\beta_0 + \mu_{01}$		0.20 (0.10)	-0.17 (0.12)	
$\beta_0 + \mu_{02}$		-0.83 (0.11)	-0.41 (0.08)	
β_1	1.03 (0.06)	1.03 (0.05)		
$\beta_1 + \mu_{11}$			2.47 (0.15)	2.64 (0.16)
$\beta_1 + \mu_{12}$			0.50 (0.10)	0.50 (0.09)
β_2	0.70 (0.05)	0.70 (0.04)	0.72 (0.04)	0.72 (0.04)
β_3	1.36 (0.05)	1.36 (0.03)	1.41 (0.03)	1.41 (0.03)
β_4	-0.67 (0.06)	-0.67 (0.04)	-0.69 (0.04)	-0.69 (0.04)
d_{11}	1.86 (0.20)	1.53 (0.16)	1.30 (0.10)	1.35 (0.10)
p_1		0.52 (0.07)	0.30 (0.05)	0.27 (0.04)
p_2		0.48 (0.01)	0.70 (0.05)	0.73 (0.04)
Log-likelihood	-4116.05	-4115.39	-4079.07	-4079.84

which assumes that at most two different values are possible for the intercepts, as well as for the slopes (row 1 in Figure 23.2). Depending on the actual location of the mean parameters $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, the model further reduces to a one-component mixture (column A in Figure 23.2), or to a two-component mixture with discrimination in only one dimension or in both dimensions (columns B and C, respectively, in Figure 23.2). A similar column-classification is also possible in case one dimension of the random-effects distribution shows within-component variability (row 2 in Figure 23.2), or when within-component variability is present in both dimensions (row 3 in Figure 23.2).

As an example, several of these models have been fitted to the verbal aggression data, all assuming within-component variability for the intercepts (i.e., $d_{11} > 0$), but a latent class structure for the effect of the behavior mode (i.e., $d_{12} = d_{22} = 0$). Hence, all models are of the type as shown in row 2 of Figure 23.2. The results have been summarized in Table 23.3. First, the homogeneity model, i.e., a one-component model, was fitted ($g = 1$). Clearly, people tend to be more verbally aggressive when others are to blame and when the considered behavior is expressing blame or expressing frustration. Moreover, they want to be more aggressive than they say they would actually be.

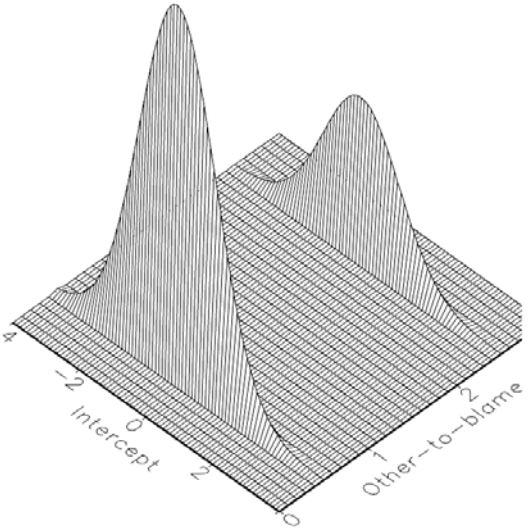

FIGURE 23.3. *Verbal Aggression Data. Fitted random-effects distribution based on the two-component mixture model, Model B.*

Our first two-component mixture model (Model A) assumes a two-component mixture for the intercepts, but still one common effect of the covariate X_1. More specifically, we assume that

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim p_1 \, N\left[\begin{pmatrix} \mu_{01} \\ 0 \end{pmatrix}, \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix}\right] + p_2 \, N\left[\begin{pmatrix} \mu_{02} \\ 0 \end{pmatrix}, \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix}\right],$$

which graphically corresponds to panel (2.B) in Figure 23.2. The two mixture components get estimated weights (prior probabilities) equal to 0.52 and 0.48. Note that the results in Table 23.3 are the component means μ_{01} and μ_{02}, with the fixed effect β_0 added, yielding the average intercept within each mixture component separately. In case β_0 would be of interest, the estimate immediately follows from the fact that

$$\beta_0 \;=\; p_1(\mu_{01} + \beta_0) \;+\; p_2(\mu_{02} + \beta_0),$$

because the random effects have been assumed to have prior mean equal to zero. In our example, this yields

$$\widehat{\beta_0} \;=\; 0.52 \times 0.20 \;-\; 0.48 \times 0.83 \;=\; -0.29,$$

relatively close to the overall intercept we obtained under the homogeneity model. Note also the reduction in within-component variability d_{11}. Finally, although classical likelihood ratio tests for the comparison of the one-component model with Model A are not valid (see Section 23.2), comparison of the log-likelihood values does yield very little evidence in favor of the two-component model.

Model A assumes the same effect of X_1 in both mixture components. In Model B, this is relaxed by assuming that

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim p_1\, N\left[\begin{pmatrix} \mu_{01} \\ \mu_{11} \end{pmatrix}, \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix}\right] + p_2\, N\left[\begin{pmatrix} \mu_{02} \\ \mu_{12} \end{pmatrix}, \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix}\right],$$

graphically represented in panel (2.C) of Figure 23.2. Clearly, this model yields an improved fit, when compared to Model A. Figure 23.3 shows the fitted random-effects distribution. The smaller class represents approximately 30% of the population, the larger class 70%. Figure 23.3 clearly shows that a major distinction between the two mixture components is given by the effect of the 'other-to-blame' factor. Our homogeneity model showed that verbal aggression is higher when others are to blame, compared to situations in which one should blame oneself. In the smaller class this difference is much higher than in the larger class (2.474 *versus* 0.501). This means that there are two types of people: Those who do not differentiate very much between other-to-blame situations and self-to-blame situations and those who are clearly more verbally aggressive when others are to blame.

Figure 23.3 also suggests that there is very little differentiation between the mixture components with respect to the random intercepts: The average intercepts in the two components are estimated as -0.167 in the first component versus -0.414 in the second mixture component. Therefore, a two-component model, with a common average random intercept for both components has also been fitted (Model C). The random effects are then assumed to satisfy

$$\begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim p_1\, N\left[\begin{pmatrix} 0 \\ \mu_{11} \end{pmatrix}, \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix}\right] + p_2\, N\left[\begin{pmatrix} 0 \\ \mu_{12} \end{pmatrix}, \begin{pmatrix} d_{11} & 0 \\ 0 & 0 \end{pmatrix}\right].$$

The maximized log-likelihood value is now -4079.84, which is only slightly smaller than what was obtained under Model B.

23.6 Concluding Remarks

In linear mixed models, inferences for the fixed effects and variance components are quite robust with respect to non-normality of the random effects. This no longer holds for non-linear or generalized linear mixed models. We have presented a flexible class of models with less strict distributional assumptions for the random effects, which includes the traditional mixed models based on Gaussian random effects, as special cases.

In the analysis of the verbal aggression data (Section 23.5), we have illustrated the flexibility of the models, in the context of a mixed logistic model for a binary outcome variable. However, the heterogeneity model can equally well be applied to non-linear mixed models (Section 20.5).

Note also that many further extensions of the models presented in the example in Section 23.5 would be possible. The number of mixture components could be further increased, class-specific variances could be assumed, within-component variability could be allowed for the effects of the type of situation, or other random effects could be included as well. Our purpose has been to illustrate the flexibility of the heterogeneity model, rather than to give a complete overview of all possible models that fit within this framework.