8

Finite Mixtures of Regression Models

8.1 Introduction

In applied statistics as well as in econometrics a tremendous amount of applications deal with relating a random variable Y_i, which is observed on several occasions $i = 1, \ldots, N$, to a set of explanatory variables or covariates $(z_{i1}, \ldots, z_{i,d-1})$ through a regression-type model, where the conditional mean of Y_i is assumed to depend on $\mathbf{x}_i = \big(z_{i1} \cdots z_{i,d-1} \; 1 \big)$ through $\mathrm{E}(Y_i|\boldsymbol{\beta}, \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of unknown regression coefficients of dimension d.

In many circumstances, however, the assumption that the regression coefficient is fixed over all possible realizations of $Y_1, \ldots, Y_N$ is inadequate, and models where the regression coefficient changes are of great practical importance. The most general alternative is to assume a different regression coefficient $\boldsymbol{\beta}_i^s$ for each realization Y_i, $\mathrm{E}(Y_i|\boldsymbol{\beta}, \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}_i^s$, however, only in rare cases will it be possible to estimate $\boldsymbol{\beta}_i^s$ without imposing further structure, and modeling $\boldsymbol{\beta}^s = (\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s)$ becomes an important issue.

For identifying a sensible model for $\boldsymbol{\beta}^s$, it is helpful to understand why the regression coefficients are different. For sequential observations the regression coefficient may change over time, whereas for cross-sectional data the regression coefficient may change between subgroups of observations. In both cases the model may be misspecified because of omitted variables and nonlinearities or the sample may contain outliers. Whatever information is available about the nature of heterogeneity for the problem at hand should be incorporated in an appropriate manner. Within a Bayesian approach, this information is included by choosing a specific probabilistic model for $\boldsymbol{\beta}^s$ which is specified in terms of the density $p(\boldsymbol{\beta}^s)$ of the joint distribution of $\boldsymbol{\beta}^s = (\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s)$. $p(\boldsymbol{\beta}^s)$ plays the role of a prior distribution, imposing some model structure on the individual regression coefficients that may be overruled by the information in the data. Different such prior distributions defining different model structures may be compared in a principled way by Bayesian model comparison.

This chapter focuses on capturing parameter heterogeneity for cross-sectional data through finite mixtures of regression models where changes in $\boldsymbol{\beta}_i^s$ are driven by a hidden discrete indicator S_i, which is allowed to take one out of K values for each observation Y_i. This model is formulated in Section 8.2, whereas statistical inference is discussed in Section 8.3.

Several useful extensions of this model are discussed in this chapter, such as mixed-effects finite mixtures of regression models in Section 8.4, which combine regression coefficients that are fixed across all realizations with regression coefficients that are allowed to change, and finite mixtures of random-effects models in Section 8.5, which are useful for longitudinal data and repeated measurements.

8.2 Finite Mixture of Multiple Regression Models

In this section focus lies on extending the standard multiple regression model with normally distributed errors by introducing a regression coefficient that changes between groups of otherwise homogeneous observations.

8.2.1 Model Definition

Let $(Y_i, \boldsymbol{z}_i)$ be a pair of a random variable Y_i and a set of explanatory variables $\boldsymbol{z}_i = (z_{i1}, \dots, z_{i,d-1})$. Suppose that dependence of Y_i on $\boldsymbol{z}_i$ is modeled by a multiple regression model:

$$Y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{8.1}$$

where $\mathbf{x}_i = \begin{pmatrix} z_{i1} & \cdots & z_{i,d-1} & 1 \end{pmatrix}$ is a design point, and $\boldsymbol{\beta}$ and σ_ε^2 are unknown parameters. Assume that background information suggests that the regression coefficient $\boldsymbol{\beta}$ and the error variance σ_ε^2 are not homogeneous over all possible pairs $(Y_i, \boldsymbol{z}_i)$. One way to capture such changes in the parameter of a regression model is finite mixtures of regression models. A finite mixture regression model assumes that a set of K regression models characterized by the parameters $(\boldsymbol{\beta}_1, \sigma_{\varepsilon,1}^2), \dots, (\boldsymbol{\beta}_K, \sigma_{\varepsilon,K}^2)$ exists, and that for each observation pair $(Y_i, \boldsymbol{z}_i)$ a hidden random indicator S_i chooses one among these models to generate Y_i:

$$Y_i = \mathbf{x}_i \boldsymbol{\beta}_{S_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_{\varepsilon,S_i}^2\right). \tag{8.2}$$

$\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$ as well as $\sigma_{\varepsilon,1}^2, \dots, \sigma_{\varepsilon,K}^2$ are unknown parameters that need to be estimated from the data. The statistician applying a finite mixture of regression models has to specify how the random mechanism S_i works. In the absence of any additional information it is usual to assume that S_i and $S_{i'}$ are pairwise independent, and each S_i is distributed according to an unknown probability distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$. In what follows, $\boldsymbol{\vartheta}$ summarizes all unknown model parameters, including the parameters $\boldsymbol{\eta}$ appearing in the definition of the distribution law of $\mathbf{S} = (S_1, \dots, S_N)$.

It is easy to verify that the marginal distribution of Y_i, when holding the design point $\mathbf{x}_i$ as well as $\boldsymbol{\vartheta}$ fixed, reads:

$$p(y_i|\mathbf{x}_i, \boldsymbol{\vartheta}) = \sum_{k=1}^{K} p(y_i|\mathbf{x}_i, S_i, \boldsymbol{\vartheta})\Pr(S_i = k|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k f_N(y_i; \mathbf{x}_i\boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k}).$$

Thus for each value of the design point $\mathbf{x}_i$, the marginal distribution of Y_i is a finite mixture of univariate normal distributions with mean $\mu_{k,i} = \mathbf{x}_i\boldsymbol{\beta}_k$ and variance $\sigma^2_{\varepsilon,k}$. Therefore a finite mixture of regression models may be seen as an extension of a finite mixture of univariate normal distributions where the mean in the mixture distribution depends on explanatory variables. On the other hand, a finite mixture of univariate normal distributions may be seen as that special case of finite mixtures of regression models where $\beta_k = \mu_k$ and $\mathbf{x}_i = 1$ for all $i = 1, \ldots, N$.

Various extensions of model (8.2) are useful. The mixture regression model defined in (8.2) is heteroscedastic because the variance of the error term ε_i changes across the realizations. If the variance of the error term is unaffected by S_i, a homoscedastic finite mixture of regression models results:

$$Y_i = \mathbf{x}_i\boldsymbol{\beta}_{S_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_\varepsilon\right). \tag{8.3}$$

The distributional law of $\mathbf{S}$ may be substituted by other structures, if more information about the nature of heterogeneity is available. As discussed in Subsection 8.6.2, the probability of belonging to a certain state may depend on a covariate. For random covariates, the covariate distribution may differ between the clusters, in which case a multivariate finite normal mixture model as discussed in Chapter 6 may be appropriate. Whenever data are collected sequentially, alternative probability structures on the hidden indicator turn out to be useful. Goldfeld and Quandt (1973) introduced a hidden Markov chain into a mixture regression model, in order to deal with time series data that depend on exogenous variables. This issue is discussed in Subsection 10.3.2.

8.2.2 Identifiability

Like any finite mixture model, finite mixtures of regression models suffer from nonidentifiability due to label switching and potential overfitting; see Section 1.3 for a general discussion of these issues. More importantly, generic identifiability of finite mixtures of regression models does not in general follow from the generic identifiability of Gaussian mixtures as falsely claimed, for instance, in DeSarbo and Cron (1988), despite the close relationship between these two model classes.

A necessary condition for identifiability of a standard regression model is that the matrix $\mathbf{X}'\mathbf{X}$, where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{pmatrix},$$

is of full rank. For finite mixtures of regression models nonidentifiability may occur, even if this condition is fulfilled. This was first noticed by Hennig (2000) who showed that the regression parameters are identifiable, iff the number K of clusters is smaller than the number of distinct $(d-1)$-dimensional hyperplanes generated by the covariates (excluding the constant). Loosely speaking, identifiability problems occur for finite mixtures of regression models with covariates that show too little variability. Problems are to be expected, in particular, if covariates are dummy variables or reflect a few categories as in marketing research. In this section we provide more details on this important issue.

Consider the set of different covariates $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$. Assume that for each covariate $\mathbf{x}_i$ that the identity

$$\sum_{k=1}^{K} \eta_k f_N(y; \mu_{k,i}, \sigma^2_{\varepsilon,k}) = \sum_{k=1}^{K} \eta^\star_k f_N(y; \mu^\star_{k,i}, \sigma^{2,\star}_{\varepsilon,k}), \tag{8.4}$$

where $\mu_{k,i} = \mathbf{x}_i \boldsymbol{\beta}_k$ and $\mu^\star_{k,i} = \mathbf{x}_i \boldsymbol{\beta}^\star_k$, holds for all $y \in \Re$. If the model parameters $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,K}, \eta_1, \ldots, \eta_K)$ and $(\boldsymbol{\beta}^\star_1, \ldots, \boldsymbol{\beta}^\star_K, \sigma^{2,\star}_{\varepsilon,1}, \ldots, \sigma^{2,\star}_{\varepsilon,K}, \eta^\star_1, \ldots, \eta^\star_K)$ are related to each other by relabeling, then the finite mixture regression model is generically identifiable.

For a fixed covariate $\mathbf{x}_i$, (8.4) reduces to a Gaussian mixture, and generic identifiability of Gaussian mixtures implies the existence of a permutation $\rho_i(\cdot)$ of $\{1, \ldots, K\}$ such that for all $k = 1, \ldots, K$:

$$\eta^\star_k = \eta_{\rho_i(k)}, \qquad \mathbf{x}_i \boldsymbol{\beta}^\star_k = \mathbf{x}_i \boldsymbol{\beta}_{\rho_i(k)}, \qquad \sigma^{2,\star}_{\varepsilon,k} = \sigma^2_{\varepsilon,\rho_i(k)}. \tag{8.5}$$

A major cause for generic nonidentifiability is that the different permutations $\rho_i(\cdot)$ appearing in (8.5) are not necessarily the same for all design points $\mathbf{x}_i$, $i = 1, \ldots, p$.

Nevertheless, let us assume for the moment that actually the same permutation $\rho_s(\cdot)$ has been applied for all design points $\mathbf{x}_i$, $i = 1, \ldots, p$. Then (8.5) implies $\mathbf{x}_i \boldsymbol{\beta}^\star_k = \mathbf{x}_i \boldsymbol{\beta}_{\rho_s(k)}$ for all $i = 1, \ldots, p$ and:

$$\mathbf{X} \boldsymbol{\beta}^\star_k = \mathbf{X} \boldsymbol{\beta}_{\rho_s(k)},$$

where the rows of the design matrix $\mathbf{X}$ are equal to $\mathbf{x}_1, \ldots, \mathbf{x}_p$. If $\mathbf{X}^{'}\mathbf{X}$ has full rank, then it follows immediately that the regression coefficients are determined up to relabeling:

$$\boldsymbol{\beta}^\star_k = \boldsymbol{\beta}_{\rho_s(k)}, \tag{8.6}$$

ensuring generic identifiability. The problem with this derivation is, that without further assumptions, the different permutations $\rho_i(\cdot)$ appearing in (8.5) are not necessarily the same for all $i = 1, \ldots, p$.

It is possible to show that these permutations are necessarily the same, if any two regression models in the mixture differ at least in η_k or $\sigma^2_{\varepsilon,k}$. Assume that (8.5) holds for two different permutations $\rho_s(\cdot)$ and $\rho_t(\cdot)$. Then $\eta_{k_1} = \eta_{k_2}$ and $\sigma^2_{\varepsilon,k_1} = \sigma^2_{\varepsilon,k_2}$ for regression model $k_1 = \rho_s(k)$ and $k_2 = \rho_t(k)$, which contradicts the assumption made above.

If η_k and $\sigma^2_{\varepsilon,k}$ are the same for at least two regression models, then it is possible that (8.5) holds for two different permutations $\rho_s(\cdot)$ and $\rho_t(\cdot)$. Assume that $\eta_{k_1} = \eta_{k_2}$ and $\sigma^2_{\varepsilon,k_1} = \sigma^2_{\varepsilon,k_2}$. Then any two permutations where $\rho_s(k_1) = \rho_t(k_2)$, $\rho_s(k_2) = \rho_t(k_1)$, and $\rho_s(l) = \rho_t(l)$, for $l \neq k_1, k_2$, fulfill (8.5). In this case generic nonidentifiability may occur, even if the matrix $\mathbf{X}^{'}\mathbf{X}$ has full rank.

Consider, for instance, a mixture of two regression models, where $\eta_1 = \eta_2$ and $\sigma^2_{\varepsilon,1} = \sigma^2_{\varepsilon,2}$. For each $i = 1, \ldots, p$, the permutation $\rho_i(\cdot)$ appearing in (8.5) is equal to one of the two possible permutations, namely the identity, $\rho_1(1) = 1$ and $\rho_1(2) = 2$, or the permutation $\rho_2(1) = 2$ and $\rho_2(2) = 1$, which interchanges the labeling. Reorder, for a given sequence of permutations, the equations in (8.5) according to the permutation applied to $k = 1$. Then:

$$\mathbf{X}_1\boldsymbol{\beta}_1^\star = \mathbf{X}_1\boldsymbol{\beta}_1, \tag{8.7}$$

$$\mathbf{X}_2\boldsymbol{\beta}_1^\star = \mathbf{X}_2\boldsymbol{\beta}_2, \tag{8.8}$$

where the rows of the design matrix $\mathbf{X}_1$ are built from all design points $\mathbf{x}_i$, where $\rho_i(1) = 1$, and the rows of the design matrix $\mathbf{X}_2$ are built from all design points $\mathbf{x}_i$, where $\rho_i(1) = 2$. If in (8.7) and (8.8) either $\mathrm{rg}(\mathbf{X}_1^{'}\mathbf{X}_1) = d$ or $\mathrm{rg}(\mathbf{X}_2^{'}\mathbf{X}_2) = d$ (or both), then (8.6) follows immediately.

Thus generic identifiability up to relabeling follows, if $\mathrm{rg}(\mathbf{X}_1^{'}\mathbf{X}_1) = d$ or $\mathrm{rg}(\mathbf{X}_2^{'}\mathbf{X}_2) = d$ holds for any partition of the set of different covariates $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ into two sets. This is essentially the same condition as the one given by Hennig (2000). Any partition that violates this condition defines an alternative solution. It follows that for $K = 2$ the minimum number of different design points is equal to $2\dim(\boldsymbol{\beta}_k) + 1$, which is sufficient to achieve identifiability, iff all subsets of size $\dim(\boldsymbol{\beta}_k)$ define a design matrix of full rank. To give an example, consider a mixture of two regression models where $\dim(\boldsymbol{\beta}_k) = 2$ where there are only two linear independent design points $\mathbf{x}_1 = (z_1\ 1)$ and $\mathbf{x}_2 = (z_2\ 1)$. A similar example appears in Hennig (2000). Evidently the partition $\{\mathbf{x}_1\} \cup \{\mathbf{x}_2\}$ violates the rank condition. Only if the two permutations in (8.5) are the same, do we obtain (8.6). However, if the two permutations in (8.5) are different, then another solution exists, which is given by

$$\mathbf{X}^\star = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \qquad \boldsymbol{\beta}_1^\star = (\mathbf{X}^\star)^{-1} \begin{pmatrix} \mathbf{x}_1\boldsymbol{\beta}_1 \\ \mathbf{x}_2\boldsymbol{\beta}_2 \end{pmatrix}, \qquad \boldsymbol{\beta}_2^\star = (\mathbf{X}^\star)^{-1} \begin{pmatrix} \mathbf{x}_1\boldsymbol{\beta}_2 \\ \mathbf{x}_2\boldsymbol{\beta}_1 \end{pmatrix}.$$

Consequently, this finite mixture regression model is generically unidentifiable. Whereas a single regression line is determined from two covariate points, for a mixture of two regressions this is not the case. Identifiability is achieved by adding a third design point $\mathbf{x}_3 = (z_3\ 1)$, with $z_3 \neq z_1, z_2$. Then any partition of the design points into two groups contains at least two different design points and the identifiability condition is fulfilled.

Identifiability of Finite Mixtures of Regression Models

Consider a mixture of K regression models, where η_k and $\sigma^2_{\varepsilon,k}$ are the same in all groups. For each $k = 1, \dots, K$, reorder the equations in (8.5) according to the permutation applied to the label k. Then:

$$\begin{aligned} \mathbf{X}_1\boldsymbol{\beta}_k^\star &= \mathbf{X}_1\boldsymbol{\beta}_1, &(8.9)\\ \mathbf{X}_2\boldsymbol{\beta}_k^\star &= \mathbf{X}_2\boldsymbol{\beta}_2, \\ &\vdots \\ \mathbf{X}_K\boldsymbol{\beta}_k^\star &= \mathbf{X}_K\boldsymbol{\beta}_K, &(8.10) \end{aligned}$$

where the rows of the design matrix $\mathbf{X}_j$ are built from all design points $\mathbf{x}_i$, where $\rho_i(k) = j$. If in (8.9) to (8.10) $\mathrm{rg}(\mathbf{X}_j^{'}\mathbf{X}_j) = d$ for at least one $j = 1, \dots, K$, then (8.6) follows immediately. Thus generic identifiability up to relabeling follows, if $\mathrm{rg}(\mathbf{X}_j^{'}\mathbf{X}_j) = d$ holds for any partition of the set of different covariates $\{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ into K subsets. This is essentially the same condition as the one given by Hennig (2000).

Any partition that violates this condition defines an alternative solution. It follows that the minimum number of different design points is equal to $K(\dim(\boldsymbol{\beta}_k) - 1) + 1$. If $p \leq K(\dim(\boldsymbol{\beta}_k) - 1)$, then evidently there exists a partition of the different design points into K groups, where each set contains at most $\dim(\boldsymbol{\beta}_k) - 1$ design points and violates the rank condition. This minimum number of design points is sufficient to achieve identifiability, iff all subsets of size $\dim(\boldsymbol{\beta}_k)$ define a design matrix of full rank.

Grün and Leisch (2004) use bootstrap methods as a diagnostic tool for revealing identifiability problems in finite mixtures of normal and nonnormal regression models.

8.2.3 Statistical Modeling Based on Finite Mixture of Regression Models

In statistical modeling finite mixtures of regression models are also known as *switching regression models* in economics (Quandt, 1972), as *latent class regression models* in marketing (DeSarbo and Cron, 1988), as *mixture-of-expert models* in the machine-learning literature (Jacobs et al., 1991), and as *mixed models* in biology (Wang et al., 1996).

The Switching Regression Model

For sequentially observed data, one source of heterogeneity is sudden changes in regression coefficients due to a structural break. A simple model to capture a sudden parameter change at a known breakpoint τ within the standard multiple regression model is the following,

$$Y_i = \begin{cases} \mathbf{x}_i\boldsymbol{\beta}_1 + \varepsilon_i, & \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,1}\right), & i < \tau, \\ \mathbf{x}_i\boldsymbol{\beta}_2 + \varepsilon_i, & \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,2}\right), & i \geq \tau. \end{cases} \tag{8.11}$$

It is useful to reparameterize model (8.11) as

$$Y_i = \mathbf{x}_i(1 - D_i)\boldsymbol{\beta}_1 + \mathbf{x}_i D_i \boldsymbol{\beta}_2 + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_i\right), \tag{8.12}$$

where $\sigma^2_i = \sigma^2_{\varepsilon,1}(1 - D_i) + \sigma^2_{\varepsilon,2}D_i$. D_i is a dummy variable, taking the value 0 for $i < \tau$ and 1 otherwise. If the breakpoint τ is known, then D_i is exogenous, and (8.12) is a regression model with heteroscedastic errors. If the exact position of the break point τ is unknown, D_i is not observable, but a latent, discrete random variable, taking the values 0 and 1 according to some unknown probability law, and (8.11) turns out to be a finite mixture of regression models, also called a switching regression model.

An early example of a switching regression model with unknown breakpoint is considered in Quandt (1958) who studies the consumption function $Y = \beta X + \alpha$, where X is the income and Y is the consumption, and assumes that other factors, that are difficult to identify, affect the parameters of the consumption function. If this critical factor is below a threshold, then $Y = \beta_1 X + \alpha_1$, otherwise $Y = \beta_2 X + \alpha_2$. In general we are not able to identify the critical variable, and what we observe is a mixture of these two regression lines. Quandt (1958) considers a single shift between the two regimes at an unknown break point, mainly to make estimation feasible under the computational limitations of the 1950s.

A particularly important extension of this work is Quandt (1972), where for the first time a probability model is introduced, to model "that nature chooses between regimes with probability η_1 and $1 - \eta_1$"(Quandt, 1972, p.306).[1] Quandt (1972) starts directly from specifying the conditional density $p(y_i|\mathbf{x}_i, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma^2_{\varepsilon,1}, \sigma^2_{\varepsilon,2}, \eta_1)$ as a mixture of two normal distributions:

$$Y_i \sim \eta_1 \mathcal{N}\left(\mathbf{x}_i\boldsymbol{\beta}_1, \sigma^2_{\varepsilon,1}\right) + (1 - \eta_1)\mathcal{N}\left(\mathbf{x}_i\boldsymbol{\beta}_2, \sigma^2_{\varepsilon,2}\right). \tag{8.13}$$

In his summarizing remarks, Quandt (1972, p.310) concludes that "A notable disadvantage of the method is that it does not allow individual observations to be identified with particular regimes." The latent variable interpretation of his important contribution, which allows clustering observations into regimes, was discovered only later.

Further applications of switching regression models in econometrics are found in Fair and Jaffee (1972) and Quandt and Ramsey (1978), who consider the relation between wage bargains and unemployment rate through a Phillips curve which is expected to switch according to low and high changes on the consumer price index.

[1] Original notation of Quandt (1972) changed.

Omitted Categorical Predictors

Mixtures of regression models arise whenever a categorical or dummy regressor is omitted. Hosmer (1974), which is an early reference in this area, considered a mixture of two regression lines with a nice application from fishery research. In commercial catches of halibut only age and length are measured, whereas the gender of the fish is unknown. For any particular age, the mean length of female fish exceeds that of male fish, and this difference increases with age. If gender were observed, length may be modeled in terms of gender g_i and age a_i in the following way.

$$Y_i = \beta_1 + a_i\beta_2 + g_i\beta_3 + g_i a_i\beta_4 + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right).$$

When coding gender as a 0/1 variable, this model may be written as

$$Y_i = \beta_{g_i,1} + a_i\beta_{g_i,2} + \varepsilon_i, \tag{8.14}$$

where $\beta_{g_i,1} = \beta_1 + g_i\beta_3$ and $\beta_{g_i,2} = \beta_2 + g_i\beta_4$. If gender is unobserved, then (8.14) is equal to a mixture of two regression models. In a scatter plot of a_i versus the observed length y_i, the observations cluster around two regression lines, one corresponding to males, the other to females. When a switching regression model is fitted to the data, then both unknown regression lines have to be reconstructed from the data.

Note that the switching slope in (8.14) is caused by interaction between the observed and the omitted categorical variable. If such an interaction is not present, then $\beta_4 = 0$ and (8.14) reduces to a regression model with a shift in the intercept only:

$$Y_i = \beta_{g_i,1} + a_i\beta_2 + \varepsilon_i.$$

Unknown Segments in the Population

Finite mixtures of regression models, introduced into marketing by DeSarbo and Cron (1988), found numerous applications in marketing research; see Wedel and DeSarbo (1993b) and Rossi et al. (2005) for a review. In marketing, consumers rate the quality of products or events. A regression model is built to describe the relation between the rating Y_i of consumer i and certain features of the product summarized in the design matrix $\mathbf{x}_i$. If unknown segments in the population are present, then the part-worths $\boldsymbol{\beta}_i^s$ of a certain consumer i depend on membership in a certain segment. If we introduce a segment indicator S_i, then the market segmentation regression model reads:

$$Y_i = \mathbf{x}_i\boldsymbol{\beta}_{S_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right). \tag{8.15}$$

Apart from estimating the regression coefficients in the different segments, the indicator S_i itself is of interest, as it allows us to assign each consumer to a certain segment k.

8.2.4 Outliers in a Regression Model

The finite mixture model discussed in Section 7.2 for dealing with outliers in univariate data sets has been extended in several ways to deal with outliers in a linear regression model; see the review article by Peña and Guttman (1993).

Box and Tiao (1968), for instance, extend the variance inflation model (7.24) in the following way.

$$Y_i \sim (1-\eta_2)\mathcal{N}\left(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2_\varepsilon\right) + \eta_2\mathcal{N}\left(\mathbf{x}_i\boldsymbol{\beta}, k\sigma^2_\varepsilon\right). \tag{8.16}$$

Model (8.16) is a regression model with switching variances, but a constant regression parameter $\boldsymbol{\beta}$. Abraham and Box (1978) extended the location shift model (7.25) to allow for outliers in a linear regression model:

$$Y_i \sim (1-\eta_2)\mathcal{N}\left(\mathbf{x}_i\boldsymbol{\beta}, \sigma^2_\varepsilon\right) + \eta_2\mathcal{N}\left(\mathbf{x}_i\boldsymbol{\beta} + k, \sigma^2_\varepsilon\right). \tag{8.17}$$

Model (8.17) allows for a switching intercept, while holding the variance fixed. Peña and Guttman (1993) show that these models are more effective in identifying outliers than methods which postulate a null model for the generation of the data with no alternative to the null model being entertained.

Various extensions to models (8.16) and (8.17) are worth mentioning. Guttman et al. (1978) combine a mixture of a normal regression models with a random-effects model to allow for a different shift for each outlier. Outlier modeling in nonnormal mixture regression models is considered in Pregibon (1981), Copas (1988), and Verdinelli and Wasserman (1991). West (1984, 1985) also studies more general scale mixtures of GLMs to deal with outliers.

8.3 Statistical Inference for Finite Mixtures of Multiple Regression Models

Parameter estimation for finite mixtures of regression models is usually based on ML estimation or Bayesian estimation, an exception being Quandt and Ramsey (1978) who used a method of moments estimator based on the moment-generating function.

8.3.1 Maximum Likelihood Estimation

Assume that N observation pairs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ are available. The appropriate likelihood function for parameter estimation for a finite mixture of an arbitrary number K of regression models was derived for the first time by Quandt (1972). This function turns out to be the following extension of the mixture likelihood of a standard finite mixture model,

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N}\left(\sum_{k=1}^{K} f_N(y_i; \mathbf{x}_i\boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k})\eta_k\right), \tag{8.18}$$

where $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K, \sigma^2_{\varepsilon,1}, \dots, \sigma^2_{\varepsilon,K}, \boldsymbol{\eta})$. In contrast to this, Fair and Jaffee (1972) consider maximization of the classification likelihood $p(\mathbf{y}|\boldsymbol{\vartheta}, \mathbf{S})$ with respect to $\boldsymbol{\vartheta}$ and $\mathbf{S}$ for jointly solving the problem of parameter estimation and estimating the unknown allocations. However, Oberhofer (1980) showed that this approach leads in general to inconsistent estimators of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$.

In Quandt (1972), the mixture regression likelihood function $p(\mathbf{y}|\boldsymbol{\vartheta})$ is maximized numerically, and considerable convergence failures are reported for repeated experiments on artificially generated data. A mixture of two regression models, for instance, where $\boldsymbol{\beta}_1 = (1, 1)$, $\boldsymbol{\beta}_2 = (0.5, 1.5)$, $\sigma^2_{\varepsilon,1} = 2$, $\sigma^2_{\varepsilon,2} = 2.5$, $\eta_1 = \eta_2 = 0.5$, $N = 60$, and $\mathbf{x}_i = (1\, x_i)$, where $x_i \sim \mathcal{U}\,[0, 40]$, leads to a failure rates of 53 percent in 30 replications, where x_i was kept fixed over the repetitions.

Later on, Hosmer (1974) realized that the problem of dealing with an unbounded likelihood function is of relevance not only for finite mixtures of normal distributions (see again Subsection 6.1.2), but also for heterogeneous mixtures of regression models, which include heterogeneous mixtures of normal distributions as a special case. Hosmer (1974) noted that any observation y_i generates a singularity in the likelihood function if $\boldsymbol{\beta}_k$ is chosen such that $y_i = \mathbf{x}_i \boldsymbol{\beta}_k$, and $\sigma^2_{\varepsilon,k}$ goes to 0. More generally, each subgroup of d observations generates a singularity in the likelihood function if $\boldsymbol{\beta}_k$ is chosen such that the regression plane provides a perfect fit to this subgroup.

Thus if the variances of a finite mixture of regression models are unconstrained, a global maximizer of the likelihood function does not exist. Nevertheless, Kiefer (1978) shows that a root of the log likelihood equations corresponding to a local maximizer in the interior of the parameter space is consistent, asymptotically normal, and efficient. In practice, however, it may be difficult to find the ML estimator numerically. An EM-type algorithm for finding the ML estimator was suggested by Hartigan (1977), whereas DeSarbo and Cron (1988) use the EM algorithm directly for this purpose.

As for mixtures of normal distributions, it is complete ignorance about the variance ratio $\sigma^2_{\varepsilon,k}/\sigma^2_{\varepsilon,l}$ that causes problems with maximum likelihood estimation, and again the Bayesian approach, discussed in the remaining subsections, is helpful in this respect, as it allows us to bound this ratio through choosing proper priors on $\sigma^2_{\varepsilon,k}$, $k = 1, \dots, K$.

8.3.2 Bayesian Inference When the Allocations Are Known

If the allocations $\mathbf{S}$ are known, then Bayesian inference reduces to Bayesian analysis of the standard regression model as discussed first in Zellner (1971); see also Raftery et al. (1997) for a more recent review.

For each group, a separate regression model with parameters $\boldsymbol{\beta}_k$ and $\sigma^2_{\varepsilon,k}$ has to be estimated from all observations that fall into that group. In matrix notation, in each group the regression model reads:

$$\mathbf{y}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k, \qquad \boldsymbol{\varepsilon}_k \sim \mathcal{N}_{N_k}\left(\mathbf{0}, \sigma^2_{\varepsilon,k} \mathbf{I}_{N_k}\right), \tag{8.19}$$

where $N_k = \#\{i : S_i = k\}$ is equal to the number of observations in group k, $\mathbf{y}_k$ is a vector containing all observations y_i with $S_i = k$, and $\mathbf{X}_k$ is the corresponding design matrix, where each line contains the regressors $\mathbf{x}_i$ corresponding to y_i. The relevant group-specific data summaries are well known from the normal equations leading to the standard OLS estimator in econometrics:

$$\mathbf{X}_k^{'}\mathbf{y}_k = \sum_{i:S_i=k} \mathbf{x}_i^{'} y_i,$$

$$\mathbf{X}_k^{'}\mathbf{X}_k = \sum_{i:S_i=k} \mathbf{x}_i^{'} \mathbf{x}_i.$$

Note that N_k as well as both group-specific data summaries depend on $\mathbf{S}$, however, as opposed to earlier chapters this dependence is not made explicit in this chapter.

Assume that observation y_i is assigned to group k, $S_i = k$. Then the contribution of y_i to the complete-data likelihood function $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma}^{\mathbf{2}}, \mathbf{S})$ is equal to

$$p(y_i|\boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k}, S_i) = \left(\frac{1}{2\pi\sigma^2_{\varepsilon,k}}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2_{\varepsilon,k}}(y_i - \mathbf{x}_i\boldsymbol{\beta}_k)^2\right).$$

The complete-data likelihood function $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma}^{\mathbf{2}}, \mathbf{S})$ has K independent factors, each carrying all information about the parameters in a certain group:

$$p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\sigma}^{\mathbf{2}}, \mathbf{S}) = \prod_{k=1}^{K} \left(\frac{1}{2\pi\sigma^2_{\varepsilon,k}}\right)^{N_k/2} \times \exp\left(-\frac{1}{2\sigma^2_{\varepsilon,k}} \sum_{i:S_i=k} (y_i - \mathbf{x}_i\boldsymbol{\beta}_k)^2\right). \tag{8.20}$$

In a Bayesian analysis each of these factors is combined with a prior. When holding the variance $\sigma^2_{\varepsilon,k}$ fixed, the complete-data likelihood function, regarded as a function of $\boldsymbol{\beta}_k$, is the kernel of a multivariate normal distribution. Under the conjugate prior $\boldsymbol{\beta}_k \sim \mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0)$, the posterior density of $\boldsymbol{\beta}_k$ given $\sigma^2_{\varepsilon,k}$ and all observations assigned to group k, is again a density from the normal distribution, $\boldsymbol{\beta}_k|\sigma^2_{\varepsilon,k}, \mathbf{S}, \mathbf{y} \sim \mathcal{N}_d(\mathbf{b}_k, \mathbf{B}_k)$, where

$$\mathbf{B}_k = (\mathbf{B}_0^{-1} + \frac{1}{\sigma^2_{\varepsilon,k}}\mathbf{X}_k^{'}\mathbf{X}_k)^{-1}, \tag{8.21}$$

$$\mathbf{b}_k = \mathbf{B}_k(\mathbf{B}_0^{-1}\mathbf{b}_0 + \frac{1}{\sigma^2_{\varepsilon,k}}\mathbf{X}_k^{'}\mathbf{y}_k). \tag{8.22}$$

When holding the regression parameter $\boldsymbol{\beta}_k$ fixed, the complete-data likelihood function, regarded as a function of $\sigma^2_{\varepsilon,k}$, is the kernel of an inverted Gamma density. Under the conjugate inverted Gamma prior $\sigma^2_{\varepsilon,k} \sim \mathcal{G}^{-1}(c_0, C_0)$, the posterior density of $\sigma^2_{\varepsilon,k}$ given $\boldsymbol{\beta}_k$ and all observations assigned to this group, is again a density from the inverted Gamma distribution, $\sigma^2_{\varepsilon,k}|\boldsymbol{\beta}_k, \mathbf{S}, \mathbf{y} \sim \mathcal{G}^{-1}(c_k, C_k)$, where

$$c_k = c_0 + \frac{N_k}{2}, \qquad C_k = C_0 + \frac{1}{2}\boldsymbol{\varepsilon}_k^{'}\boldsymbol{\varepsilon}_k, \tag{8.23}$$

where $\boldsymbol{\varepsilon}_k = \mathbf{y}_k - \mathbf{X}_k\boldsymbol{\beta}_k$.

If both $\boldsymbol{\beta}_k$ and $\sigma^2_{\varepsilon,k}$ are unknown, a closed-form solution for the joint posterior $p(\boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k}|\mathbf{S}, \mathbf{y})$ exists only if the prior of $\boldsymbol{\beta}_k$ is restricted by assuming that the prior covariance matrix depends on $\sigma^2_{\varepsilon,k}$ through $\mathbf{B}_{0,k} = \sigma^2_{\varepsilon,k}\tilde{\mathbf{B}}_0$. Then the joint posterior factors as $p(\boldsymbol{\beta}_k|\sigma^2_{\varepsilon,k}, \mathbf{y}, \mathbf{S})p(\sigma^2_{\varepsilon,k}|\mathbf{y}, \mathbf{S})$, where density of $\boldsymbol{\beta}_k$ given $\sigma^2_{\varepsilon,k}$ arises from an $\mathcal{N}_d(\mathbf{b}_k, \mathbf{B}_k)$ distribution with

$$\mathbf{B}_k = \sigma^2_{\varepsilon,k}\tilde{\mathbf{B}}_k, \qquad \tilde{\mathbf{B}}_k = (\tilde{\mathbf{B}}_0^{-1} + \mathbf{X}_k^{'}\mathbf{X}_k)^{-1}, \tag{8.24}$$

$$\mathbf{b}_k = \tilde{\mathbf{B}}_k(\tilde{\mathbf{B}}_0^{-1}\mathbf{b}_0 + \mathbf{X}_k^{'}\mathbf{y}_k), \tag{8.25}$$

whereas the marginal posterior of $\sigma^2_{\varepsilon,k}$ is a $\mathcal{G}^{-1}(c_k, C_k)$-distribution, where c_k is the same as in (8.23), however,

$$C_k = C_0 + \frac{1}{2}\left(\mathbf{y}_k^{'}\mathbf{y}_k + \mathbf{b}_0^{'}\tilde{\mathbf{B}}_0^{-1}\mathbf{b}_0 - \mathbf{b}_k^{'}\tilde{\mathbf{B}}_k^{-1}\mathbf{b}_k\right). \tag{8.26}$$

8.3.3 Choosing Prior Distributions

The investigations of the previous subsection suggest choosing the following prior distributions for finite mixtures of regression models when the allocations are unknown, which were applied, for instance, in Hurn et al. (2003).

As a prior for the regression coefficient $\boldsymbol{\beta}_k$ one may use a conditionally conjugate prior:

$$\boldsymbol{\beta}_k|\sigma^2_{\varepsilon,k} \sim \mathcal{N}_d\left(\mathbf{b}_0, \sigma^2_{\varepsilon,k}\tilde{\mathbf{B}}_0\right), \tag{8.27}$$

which introduced prior dependence between $\boldsymbol{\beta}_k$ and $\sigma^2_{\varepsilon,k}$. Alternatively, a prior may be used, where $\boldsymbol{\beta}_k$ and $\sigma^2_{\varepsilon,k}$ are independent a priori:

$$\boldsymbol{\beta}_k \sim \mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0). \tag{8.28}$$

In both cases, the prior on $\sigma^2_{\varepsilon,k}$ is inverse Gamma, $\sigma^2_{\varepsilon,k} \sim \mathcal{G}^{-1}(c_0, C_0)$. As with for finite mixtures of normal distributions, C_0 may be considered as an unknown hyperparameter with a prior of its own, $C_0 \sim \mathcal{G}(g_0, G_0)$, in which case the resulting prior is called a hierarchical prior. The prior on the group sizes is the standard Dirichlet prior, $\boldsymbol{\eta} \sim \mathcal{D}(e_0, \ldots, e_0)$.

8.3.4 Bayesian Inference When the Allocations Are Unknown

MCMC estimation is usually carried out using data augmentation and Gibbs sampling, exceptions being Chen and Liu (1996) who discuss MCMC estimation of the allocations $\mathbf{S}$ without parameter estimation and Hurn et al. (2003) who discuss direct parameter estimation without data augmentation using the Metropolis–Hastings algorithm.

Albert and Chib (1993) consider Bayesian estimation using data augmentation and Gibbs sampling for the more general Markov mixture of regression model, however, their algorithm is also relevant for finite mixtures of regression models. They show that MCMC estimation along the lines indicated by *Algorithm 3.4* is feasible after introducing the group indicator S_i for each observation pair $(\mathbf{x}_i, y_i)$ as missing data. Justel and Peña (1996) use a similar method and show that a false convergence of the Gibbs sampler may occur when one of the groups has a much smaller variance than the other. Otter et al. (2002) consider a Bayesian approach for more general finite mixtures of multivariate regression models and discuss an application in marketing. The following algorithm provides details for finite mixtures of heteroscedastic regression models.

Algorithm 8.1: Unconstrained MCMC for a Multiple Normal Mixture Regression Model Full conditional Gibbs sampling is carried out in two steps.

(a) Parameter simulation conditional on the allocations $\mathbf{S}$:
   (a1) Sample $\boldsymbol{\eta}$ from the conditional Dirichlet posterior $p(\boldsymbol{\eta}|\mathbf{S})$ as in *Algorithm 3.4*.
   (a2) Sample each regression coefficient $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$, from the posterior distribution $\boldsymbol{\beta}_k|\sigma^2_{\varepsilon,k}, \mathbf{S}, \mathbf{y} \sim \mathcal{N}_d\left(\mathbf{b}_k, \mathbf{B}_k\right)$.
   (a3) Sample each variance $\sigma^2_{\varepsilon,k}$, $k = 1, \ldots, K$, from the posterior distribution $\sigma^2_{\varepsilon,k}|\boldsymbol{\beta}_k, \mathbf{S}, \mathbf{y} \sim \mathcal{G}^{-1}\left(c_k, C_k\right)$.

(b) Classification of each observation pair $(y_i, \mathbf{x}_i)$ conditional on $\boldsymbol{\vartheta}$: sample each element S_i of $\mathbf{S}$ from the conditional posterior $p(S_i|\boldsymbol{\vartheta}, \mathbf{y})$ given by

$$\Pr(S_i = k|\boldsymbol{\vartheta}, \mathbf{y}) \propto \eta_k f_N(y_i; \mathbf{x}_i\boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k}). \tag{8.29}$$

In step (a2), the posterior moments $\mathbf{b}_k$ and $\mathbf{B}_k$ are given by (8.21) and (8.22), whereas in step (a3) the posterior moments c_k and C_k are available from (8.23). These formulae could be applied for any prior. Under the conditionally conjugate prior or the hierarchical conditionally conjugate prior, computation of $\mathbf{b}_k$ and $\mathbf{B}_k$ may be simplified as in (8.24) and (8.25). Furthermore, under this prior, sampling of $\sigma^2_{\varepsilon,k}$ is possible from the marginal inverted Gamma posterior distribution $p(\sigma^2_{\varepsilon,k}|\mathbf{S}, \mathbf{y})$, where c_k is the same as in (8.23) and C_k is given by (8.26).

Under a hierarchical prior, where C_0 is a random hyperparameter with a prior of its own, $C_0 \sim \mathcal{G}\left(g_0, G_0\right)$, an additional step has to be added in

Algorithm 8.1 to sample C_0 from $p(C_0|\mathbf{S},\boldsymbol{\beta},\boldsymbol{\sigma^2},\mathbf{y})$, which is given by Bayes' theorem as $C_0|\mathbf{S},\boldsymbol{\beta},\boldsymbol{\sigma^2},\mathbf{y} \sim \mathcal{G}\left(g_N, G_N\right)$, where:

$$g_N = g_0 + \; Kc_0, \qquad G_N = G_0 + \sum_{k=1}^{K} \frac{1}{\sigma^2_{\varepsilon,k}}.$$

MCMC for Homoscedastic Mixtures of Regression Models

Algorithm 8.1 could be applied for Bayesian estimation of a homoscedastic finite mixture regression model, where $\sigma^2_{\varepsilon,1} = \cdots = \sigma^2_{\varepsilon,K} = \sigma^2_{\varepsilon}$, however, step (a3) has to be modified by sampling σ^2_{ε} from the appropriate posterior distribution. Under the inverted Gamma prior distribution $\sigma^2_{\varepsilon} \sim \mathcal{G}^{-1}\left(c_0, C_0\right)$, the posterior distribution is again inverted Gamma, $\sigma^2_{\varepsilon}|\boldsymbol{\beta},\boldsymbol{\sigma^2},\mathbf{S},\mathbf{y} \sim \mathcal{G}^{-1}\left(c_N, C_N\right)$, where

$$c_N = c_0 + \frac{N}{2}, \qquad C_N = C_0 + \frac{1}{2}\sum_{i=1}^{N}(y_i - \mathbf{x}_i\boldsymbol{\beta}_{S_i})^2. \tag{8.30}$$

Under the conditionally conjugate prior (8.27) on $\boldsymbol{\beta}$, it is possible to sample σ^2_{ε} from the marginal posterior $p(\sigma^2_{\varepsilon}|\boldsymbol{\sigma^2},\mathbf{S},\mathbf{y})$, where $\boldsymbol{\beta}$ is integrated out, as this density is available in closed form: $\sigma^2_{\varepsilon}|\boldsymbol{\sigma^2},\mathbf{S},\mathbf{y} \sim \mathcal{G}\left(c_N, C_N\right)$, with c_N being the same as in (8.30), whereas C_N is given by

$$C_N = C_0 + \frac{1}{2}\mathbf{b}_0^{'}\tilde{\mathbf{B}}_0^{-1}\mathbf{b}_0 + \frac{1}{2}\sum_{k=1}^{K}\left(\mathbf{y}_k^{'}\mathbf{y}_k - \mathbf{b}_k^{'}\tilde{\mathbf{B}}_k^{-1}\mathbf{b}_k\right).$$

Starting Values

Justel and Peña (1996) realized that for a finite mixture of regression models Gibbs sampling may be sensitive to choosing an appropriate initial classification. In particular under the presence of outliers that mask or swamp other observations, an erroneous initial classification of the observations will lead the algorithm to a wrong solution for thousands of iterations. As a remedy, Justel and Peña (2001) avoid random initial classification and search for a more sensible classification. They use an estimate of the covariance matrix of the allocations $\mathbf{S}$ and show that the eigenvectors associated with the nonzero eigenvalues provide information about which observations are possible outliers. The examples in Justel and Peña (2001) indicate considerable improvement of the Gibbs sampler based on these elaborated starting values.

8.3.5 Bayesian Inference Using Posterior Draws

As for a standard finite mixture model, label switching as discussed in detail in Subsection 3.5.5 is also an issue for finite mixtures of regression models.

Hurn et al. (2003) use the approach of Celeux et al. (2000) to deal with the labeling problem, by choosing that parameter for estimation which minimizes the symmetrized Kullback–Leibler distance measure, which is invariant to relabeling.

As noted by Hurn et al. (2003), a functional that is invariant to relabeling is the estimated regression hyperplane,

$$\mathrm{E}(Y_i|\mathbf{x}_i) = \sum_{k=1}^{K} \eta_k \mathbf{x}_i \boldsymbol{\beta}_k,$$

which reduces to the regression line

$$\mathrm{E}(Y_i|x_i) = \sum_{k=1}^{K} \eta_k (x_i \beta_{k,1} + \beta_{k,2})$$

for simple regression problems. In the latter case, the regression line may be visualized by showing for each MCMC draw several points from this regression line for selected values of x_i (either sampled randomly from $[x_{\mathrm{min}}, x_{\mathrm{max}}]$ for continuous covariates, or sampled randomly from the set of observed covariates).

Finding identifiability constraints is not trivial, particularly in higher dimensions, however, producing scatter plots of $\beta_{k,j}$ against $\beta_{k',j'}$ for all pairs of coefficients of $\boldsymbol{\beta}$ may be helpful, as shown, for instance, in Frühwirth-Schnatter and Kaufmann (2006a). The predicted points on the regression line could also help to identify groups. If for a certain $\mathbf{x}_i$, all simulated points obey $\mathbf{x}_i\boldsymbol{\beta}_1 < \cdots < \mathbf{x}_i\boldsymbol{\beta}_K$, then this constraint could be used for identification. Thus for a switching regression model constraints need not be simple order constraints on the regression parameter, but could also be linear constraints as applied, for instance, in Otter et al. (2002).

8.3.6 Dealing with Model Specification Uncertainty

Testing for the presence of switching regression parameters was already considered by Quandt (1958), who performed an F-Test involving the ratio of variances under a switching and a nonswitching regression model, and by Quandt (1960) who considered a likelihood ratio test.

Bayes factors for testing a switching regression model with $K = 2$ against homogeneity are considered by Peña and Tiao (1992) who investigate the relation between the Bayes factor and the Chow test introduced by Chow (1960). Otter et al. (2002) and Frühwirth-Schnatter et al. (2004) use the bridge sampling estimator of the marginal likelihoods (see also Subsection 5.4.6 for more detail on this estimator) to select the number of groups in mixtures of regression models.

Hurn et al. (2003) use the birth and death process method of Stephens (2000a), discussed in Subsection 5.2.3 in detail, to select the number of groups in a finite mixture regression model.

8.4 Mixed-Effects Finite Mixtures of Regression Models

A mixed-effects model allows us to combine regression coefficients that are fixed across all realizations $(Y_i, \mathbf{x}_i)$ with regression coefficients that are allowed to change.

8.4.1 Model Definition

A mixed-effects finite mixture of regression models results if only some regression coefficients are different among the hidden groups:

$$Y_i = \mathbf{x}_i^f \boldsymbol{\alpha} + \mathbf{x}_i^r \boldsymbol{\beta}_{S_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_i}\right), \tag{8.31}$$

where $\mathbf{x}_i^f$ are the fixed effects, whereas $\mathbf{x}_i^r$ are the random effects. A necessary condition for identifiability is that the columns of the design matrix defined by

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^f & \mathbf{x}_1^r \\ \vdots & \vdots \\ \mathbf{x}_N^f & \mathbf{x}_N^r \end{pmatrix}$$

are linearly independent.

Considering certain effects as being fixed may help to avoid generic identifiability, in particular for categorical covariates. For a regression model, where only the intercept is switching,

$$Y_i = \mathbf{x}_i \boldsymbol{\alpha} + \beta_{S_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_i}\right), \tag{8.32}$$

generic identifiability follows immediately from pointwise identifiability, given by (8.5):

$$\eta_k^\star = \eta_{\rho_i(k)}, \qquad \beta_k^\star + \mathbf{x}_i \boldsymbol{\alpha} = \beta_{\rho_i(k)} + \mathbf{x}_i \boldsymbol{\alpha}, \qquad \sigma^{2,\star}_{\varepsilon,k} = \sigma^2_{\varepsilon,\rho_i(k)},$$

hence $\beta_k^\star = \beta_{\rho_i(k)}$. For the general mixed-effects model defined in (8.31) pointwise identifiability, given by (8.5),

$$\eta_k^\star = \eta_{\rho_i(k)}, \qquad \mathbf{x}_i^f \boldsymbol{\alpha} + \mathbf{x}_i^r \boldsymbol{\beta}_k = \mathbf{x}_i^f \boldsymbol{\alpha} + \mathbf{x}_i^r \boldsymbol{\beta}_{\rho_i(k)}, \qquad \sigma^{2,\star}_{\varepsilon,k} = \sigma^2_{\varepsilon,\rho_i(k)}, \tag{8.33}$$

implies $\mathbf{x}_i^r \boldsymbol{\beta}_k = \mathbf{x}_i^r \boldsymbol{\beta}_{\rho_i(k)}$, and generic identifiability holds if the identifiability condition discussed in Section 8.2.2 is applied to the design points defining only the random effects $\mathbf{x}_i^r$.

8.4.2 Choosing Priors for Bayesian Estimation

It is assumed that the priors of all parameters but $\boldsymbol{\alpha}$ are the same as in Subsection 8.3.3, whereas $\boldsymbol{\alpha} \sim \mathcal{N}_r(\mathbf{a}_0, \mathbf{A}_0)$. If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_k$ are pairwise independent a priori, then the joint prior on $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ is a normal prior, $\boldsymbol{\alpha}^* \sim \mathcal{N}_{r^*}(\mathbf{a}_0^*, \mathbf{A}_0^*)$, where $r^* = r + Kd$ and $\mathbf{a}_0^*$ and $\mathbf{A}_0^*$ are derived from $\mathbf{a}_0, \mathbf{A}_0$, $\mathbf{b}_0$, and $\mathbf{B}_0$ in an obvious way.

8.4.3 Bayesian Parameter Estimation When the Allocations Are Known

In matrix notation, in each group the regression model reads:

$$\mathbf{y}_k = \mathbf{X}_k^f \boldsymbol{\alpha} + \mathbf{X}_k^r \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k, \qquad \boldsymbol{\varepsilon}_k \sim \mathcal{N}_{N_k}\left(\mathbf{0}, \sigma^2_{\varepsilon,k}\mathbf{I}_{N_k}\right),$$

where $N_k = \#\{i : S_i = k\}$ is equal to the number of observations in group k, $\mathbf{y}_k$ is a vector containing all observations y_i with $S_i = k$, and $\mathbf{X}_k^f$ and $\mathbf{X}_k^r$ are the corresponding design matrices, where each line contains the regressors $\mathbf{x}_i^f$ and $\mathbf{x}_i^r$ corresponding to y_i.

Due to the presence of the common regression parameter $\boldsymbol{\alpha}$ in each group, conditional independence across the groups as in Subsection 8.3.2 is lost, even conditional on known allocations $\mathbf{S}$, and inference is carried out simultaneously for all regression coefficients $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$. This inference problem is closely related to Bayesian inference for a single regression model. By introducing a dummy coding for S_i through K binary variables $D_{ik}, k = 1, \dots, K$, where $D_{ik} = 1$, iff $S_i = k$, and 0 otherwise, model (8.31) is written as a heteroscedastic regression model with regression parameter $\boldsymbol{\alpha}^*$:

$$\begin{gathered} y_i = \mathbf{x}_i^f \boldsymbol{\alpha} + \mathbf{x}_i^r D_{i1} \boldsymbol{\beta}_1 + \cdots + \mathbf{x}_i^r D_{iK} \boldsymbol{\beta}_K + \varepsilon_i, \\ \varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right), \qquad \sigma_i^2 = D_{i1}\sigma^2_{\varepsilon,1} + \cdots + D_{iK}\sigma^2_{\varepsilon,K}. \end{gathered} \tag{8.34}$$

Normalization yields a regression model with homoscedastic errors:

$$\frac{y_i}{\sigma_i} = \frac{1}{\sigma_i}\mathbf{x}_i^f \boldsymbol{\alpha} + \frac{1}{\sigma_i}\mathbf{x}_i^r D_{i1}\boldsymbol{\beta}_1 + \cdots + \frac{1}{\sigma_i}\mathbf{x}_i^r D_{iK}\boldsymbol{\beta}_K + \tilde{\varepsilon}_i, \tag{8.35}$$

where $\tilde{\varepsilon}_i \sim \mathcal{N}(0,1)$. Under a normal prior on the regression coefficients $\boldsymbol{\alpha}^*$, $\boldsymbol{\alpha}^* \sim \mathcal{N}_{r^*}(\mathbf{a}_0^*, \mathbf{A}_0^*)$, the joint posterior of $\boldsymbol{\alpha}^*$, conditional on knowing the variance parameters $\sigma^2_{\varepsilon,1}, \dots, \sigma^2_{\varepsilon,K}$, is again a normal distribution: $\boldsymbol{\alpha}^* | \sigma^2_{\varepsilon,1}, \dots, \sigma^2_{\varepsilon,K}, \mathbf{y}, \mathbf{S} \sim \mathcal{N}_{r^*}(\mathbf{a}_N^*, \mathbf{A}_N^*)$. $\mathbf{a}_N^*$ and $\mathbf{A}_N^*$ are given by:

$$(\mathbf{A}_N^*)^{-1} = (\mathbf{A}_0^*)^{-1} + \sum_{i=1}^N \frac{1}{\sigma^2_{\varepsilon,S_i}} \boldsymbol{Z}_i^{'} \boldsymbol{Z}_i, \tag{8.36}$$

$$\mathbf{a}_N^* = \mathbf{A}_N^* \left((\mathbf{A}_0^*)^{-1}\mathbf{a}_0^* + \sum_{i=1}^N \frac{1}{\sigma^2_{\varepsilon,S_i}} \boldsymbol{Z}_i^{'} y_i \right), \tag{8.37}$$

where $\boldsymbol{Z}_i = (\mathbf{x}_i^f \;\; \mathbf{x}_i^r D_{i1} \;\; \cdots \;\; \mathbf{x}_i^r D_{iK})$. If N is not too large, these moments could be determined from a single matrix manipulation:

$$\begin{gathered} (\mathbf{A}_N^*)^{-1} = (\mathbf{A}_0^*)^{-1} + \mathbf{X}^{'}\mathbf{X} \\ \mathbf{a}_N^* = \mathbf{A}_N^* \left((\mathbf{A}_0^*)^{-1}\mathbf{a}_0^* + \mathbf{X}^{'}\tilde{\mathbf{y}} \right), \end{gathered}$$

where

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{Z}_1/\sigma_{\varepsilon,S_1} \\ \vdots \\ \boldsymbol{Z}_N/\sigma_{\varepsilon,S_N} \end{pmatrix}, \qquad \tilde{\mathbf{y}} = \begin{pmatrix} y_1/\sigma_{\varepsilon,S_1} \\ \vdots \\ y_N/\sigma_{\varepsilon,S_N} \end{pmatrix}.$$

In contrast to the regression parameters, the variance parameters $\sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,K}$ are independent, conditional on knowing $\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. Under the conjugate inverted Gamma prior $\sigma^2_{\varepsilon,k} \sim \mathcal{G}^{-1}(c_0, C_0)$, the posterior density of $\sigma^2_{\varepsilon,k}$ given $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_k$, and all observations assigned to this group, is again a density from the inverted Gamma distribution, $\sigma^2_{\varepsilon,k}|\boldsymbol{\alpha}, \boldsymbol{\beta}_k, \mathbf{S}, \mathbf{y} \sim \mathcal{G}^{-1}(c_k, C_k)$, where

$$c_k = c_0 + \frac{N_k}{2}, \qquad C_k = C_0 + \frac{1}{2}\boldsymbol{\varepsilon}_k^{'}\boldsymbol{\varepsilon}_k, \tag{8.38}$$

where $\boldsymbol{\varepsilon}_k = \mathbf{y}_k - \mathbf{X}_k^f\boldsymbol{\alpha} - \mathbf{X}_k^r\boldsymbol{\beta}_k$.

8.4.4 Bayesian Parameter Estimation When the Allocations Are Unknown

Bayesian parameter estimation using data augmentation and MCMC as in *Algorithm 8.1* is easily adapted to deal with mixed-effects finite mixtures of regression models.

Algorithm 8.2: Unconstrained MCMC for a Mixed-Effects Normal Mixture Regression Model Full conditional Gibbs sampling is carried out in two steps.

(a) Parameter simulation conditional on the allocations $\mathbf{S}$:
  (a1) Sample $\boldsymbol{\eta}$ from the conditional Dirichlet posterior $p(\boldsymbol{\eta}|\mathbf{S})$ as in *Algorithm 3.4*.
  (a2) Sample all regression coefficients $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ jointly from the posterior distribution $\boldsymbol{\alpha}^*|\sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,K}, \mathbf{y}, \mathbf{S} \sim \mathcal{N}_{r^\star}(\mathbf{a}_N^*, \mathbf{A}_N^*)$.
  (a3) Sample each variance $\sigma^2_{\varepsilon,k}$, $k = 1, \ldots, K$, from the posterior distribution $\sigma^2_{\varepsilon,k}|\boldsymbol{\alpha}, \boldsymbol{\beta}_k, \mathbf{S}, \mathbf{y} \sim \mathcal{G}^{-1}(c_k, C_k)$.

(b) Classification of each observation $(y_i, \mathbf{x}_i)$ conditional on $\boldsymbol{\vartheta}$: sample each element S_i of $\mathbf{S}$ from the conditional posterior $p(S_i|\boldsymbol{\vartheta}, \mathbf{y})$ given by

$$\Pr(S_i = k|\boldsymbol{\vartheta}, \mathbf{y}) \propto \eta_k f_N(y_i; \mathbf{x}_i^f\boldsymbol{\alpha} + \mathbf{x}_i^r\boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k}). \tag{8.39}$$

In step (a3), the posterior moments c_k and C_k are available from (8.38). In step (a2), joint sampling of all regression parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ is easily carried out from the conditional posterior $\mathcal{N}_{r^\star}(\mathbf{a}_N^*, \mathbf{A}_N^*)$, where the moments are given by (8.36) and (8.37). With increasing number K of groups joint sampling may be rather timeconsuming, especially for regression models with high-dimensional parameter vectors. Then one of the following variants may be useful.

Variants of Sampling the Regression Parameters for a Mixed-Effects Model

As $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are independent conditional on $\boldsymbol{\alpha}$, sampling in step (a2) of *Algorithm 8.2* may be carried out in two subblocks as in Albert and Chib (1993):

(a2-1) Conditional on $\boldsymbol{\alpha}$, sample $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ independently for each group from the regression model:

$$\mathbf{y}_k - \mathbf{X}_k^f \boldsymbol{\alpha} = \mathbf{X}_k^r \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k, \qquad \boldsymbol{\varepsilon}_k \sim \mathcal{N}_{N_k}\left(\mathbf{0}, \sigma^2_{\varepsilon,k} \mathbf{I}_{N_k}\right),$$

where only observations with $S_i = k$ are considered. This is exactly the same situation as in Subsection 8.3.2, with a slight modification of the left-hand side variable.

(a2-2) Conditional on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, sample $\boldsymbol{\alpha}$ from the posterior obtained from the regression model:

$$y_i - \mathbf{x}_i^r \boldsymbol{\beta}_{S_i} = \mathbf{x}_i^f \boldsymbol{\alpha} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_i}\right),$$

where $i = 1, \ldots, N$.

This sampler may be less efficient than joint sampling of all regression coefficients as in step (a2) of *Algorithm 8.2*, in particular if posterior correlations are high among parameters appearing in different blocks.

The following variant which has been suggested by Frühwirth-Schnatter et al. (2004) is equivalent to joint sampling of all parameters as in step (a2) of *Algorithm 8.2* and is based on decomposing the joint posterior as

$$p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\alpha}|\mathbf{S}, \sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,K}, \mathbf{y}) = \prod_{k=1}^K p(\boldsymbol{\beta}_k|\mathbf{S}, \sigma^2_{\varepsilon,k}, \mathbf{y}) p(\boldsymbol{\alpha}|\mathbf{S}, \sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,K}, \mathbf{y}).$$

The group-specific parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are sampled conditional on $\boldsymbol{\alpha}$ as in step (a2-1) above. To sample $\boldsymbol{\alpha}$, however, the marginal posterior density $p(\boldsymbol{\alpha}|\mathbf{S}, \sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,K}, \mathbf{y})$ is considered. The moments of this density are derived in Frühwirth-Schnatter et al. (2004).

8.5 Finite Mixture Models for Repeated Measurements

An often occurring problem in applied statistics is simultaneous inference on a set of parameters for similar units such as schools from a certain region, firms from the same branch, or consumers in a market. In economics, for instance, data may be available for many countries for several years, whereas in marketing the purchase behavior of many consumers may be observed on several occasions. In econometrics such data are referred to as panel data (Baltagi, 1995), whereas in statistics they are more commonly called longitudinal data (Verbeke and Molenberghs, 2000) or repeated measurements (Crowder and Hand, 1990; Davidian and Giltinan, 1998). In this section we discuss some finite mixture models that are useful for such data.

8.5.1 Pooling Information Across Similar Units

Assume that for N units i, $i = 1, \dots, N$, outcomes y_{it} are observed on several occasions $t = 1, \dots T_i$ where T_i may vary between units. In each unit i, the outcomes y_{it} are assumed to be generated by a probability law $p(y_{it}|\boldsymbol{\beta}_i^s)$ that is governed by a unit-specific parameter $\boldsymbol{\beta}_i^s$ of dimension d. It is to be expected that the parameters $\boldsymbol{\beta}_1^s, \dots, \boldsymbol{\beta}_N^s$ albeit being different across the units are related to each other. One way to model such a relation is to assume that $\boldsymbol{\beta}_i^s$ is drawn from some distribution $p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta})$ which may depend on some unknown hyperparameter $\boldsymbol{\vartheta}$. Note, however, that the distribution $p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta})$ is unknown and needs to be estimated from the data. This problem is known as unobserved heterogeneity in marketing and economics, as residual heterogeneity in the social sciences, and as frailty in medical statistics.

One way to capture unobserved heterogeneity is to assume the existence of K subpopulations of size $\eta_1, \dots, \eta_K$ with $\boldsymbol{\beta}_i^s$ being equal to a group-specific parameter $\boldsymbol{\beta}_k$ within subpopulation k. The distribution $p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta})$ is a discrete distribution with K unknown support points $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K$, where $\Pr(\boldsymbol{\beta}_i^s = \boldsymbol{\beta}_k) = \eta_k$. Alternatively, it is common to assume random deviation of $\boldsymbol{\beta}_i^s$ from a population mean $\boldsymbol{\beta}$ following a normal distribution, $\boldsymbol{\beta}_i^s \sim \mathcal{N}_d(\boldsymbol{\beta}, \mathbf{Q})$, with $\boldsymbol{\beta}$ and $\mathbf{Q}$ being unknown parameters. Without much thought the normality assumption is almost automatically taken for granted, however, as shown by Heckman and Singer (1984), the distribution of heterogeneity is rather influential and quite small changes may lead to substantial changes in the estimated parameters. The effect of misspecifying the distribution of heterogeneity is also discussed in Verbeke and Lesaffre (1997).

To achieve some robustness against the misspecification of this distribution, West (1985) chooses Student-t distributions of heterogeneity instead of normal ones, whereas Verbeke and Lesaffre (1996) choose a mixture of multivariate normal distributions to capture unobserved heterogeneity:

$$\boldsymbol{\beta}_i^s \sim \sum_{k=1}^{K} \eta_k \mathcal{N}_d(\boldsymbol{\beta}_k, \mathbf{Q}_k).$$

This distribution of heterogeneity has been called shrinkage within clusters by Frühwirth-Schnatter and Kaufmann (2006b).

8.5.2 Finite Mixtures of Random-Effects Models

The linear mixed-effects model for modeling longitudinal data was introduced by Laird and Ware (1982) and reads for each unit i:

$$y_{it} = \mathbf{x}_{it}^f \boldsymbol{\alpha} + \mathbf{x}_{it}^r \boldsymbol{\beta}_i^s + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{8.40}$$

for $t = 1, \ldots, T_i$. $\mathbf{x}_{it}^f$ is the $(1 \times r)$ design matrix for the unknown coefficient $\boldsymbol{\alpha}$, where $r = \dim(\boldsymbol{\alpha})$. $\mathbf{x}_{it}^r$ is a $(1 \times d)$ design matrix for the unknown coefficient $\boldsymbol{\beta}_i^s$, where $d = \dim(\boldsymbol{\beta}_i^s)$. $\mathbf{x}_{it}^f$ are called the fixed effects, because changing $\mathbf{x}_{it}^f$ by the same $(1 \times r)$ vector Δ changes the mean of y_{it} by the same constant $\Delta\boldsymbol{\alpha}$ for all units i. $\mathbf{x}_{it}^r$ are called the random effects, because changing $\mathbf{x}_{it}^r$ by the same $(1 \times d)$ vector Δ changes the mean of y_{it} by $\Delta\boldsymbol{\beta}_i^s$, which is different across units. Textbooks dealing with this model are Baltagi (1995), Verbeke and Molenberghs (2000), and Diggle et al. (2002).

In the standard mixed-effects model the errors ε_{it} are assumed to be homogeneous across the units. To deal with unit-specific variance heterogeneity, model (8.40) has been extended in the following way,

$$y_{it} = \mathbf{x}_{it}^f \boldsymbol{\alpha} + \mathbf{x}_{it}^r \boldsymbol{\beta}_i^s + \varepsilon_{it}, \qquad \varepsilon_{it} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2/\omega_i\right), \tag{8.41}$$

which reduces to (8.40), if $\omega_i \equiv 1$ for all $i = 1, \ldots, N$. Unit-specific scaling factors ω_i different from 1 are included to capture variance heterogeneity across the units. Like the unit-specific regression coefficients $\boldsymbol{\beta}_i^s$, the scaling factors are also assumed to arise from some distribution of variance heterogeneity, a common choice being a Gamma distribution:

$$\omega_i \sim \mathcal{G}\left(\nu/2, \nu/2\right). \tag{8.42}$$

For a fixed unit i, model (8.41) could be written as a multivariate regression model,

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_i^s + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \sigma_\varepsilon^2/\omega_i \mathbf{I}_{T_i}\right), \tag{8.43}$$

with regression parameter $(\boldsymbol{\alpha}, \boldsymbol{\beta}_i^s)$ using the matrix notation

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{i,T_i} \end{pmatrix}, \qquad \mathbf{X}_i^f = \begin{pmatrix} \mathbf{x}_{i1}^f \\ \vdots \\ \mathbf{x}_{i,T_i}^f \end{pmatrix}, \qquad \mathbf{X}_i^r = \begin{pmatrix} \mathbf{x}_{i1}^r \\ \vdots \\ \mathbf{x}_{i,T_i}^r \end{pmatrix}.$$

Note that unit-specific variances introduced through the variance model (8.42) imply the following marginal distribution for $\mathbf{y}_i$,

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_i^s + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim t_\nu\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_{T_i}\right). \tag{8.44}$$

Unobserved heterogeneity caused by omitted variables may be summarized by a regression intercept α_i that varies between the units:

$$\mathbf{y}_i = \mathbf{1}_{T_i} \alpha_i + \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \sigma_\varepsilon^2/\omega_i \mathbf{I}_{T_i}\right);$$

in other cases it will make sense to assume that all effects are random, in which case the random coefficient model results:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i^s + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \sigma_\varepsilon^2/\omega_i\mathbf{I}_{T_i}\right).$$

If $T_i \geq d$ and $\sum T_i \geq r+Kd$, then it would be possible to combine the information from all units to estimate one large regression vector $\boldsymbol{\alpha}, \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s$ without imposing further assumptions. This so-called fixed-effects approach estimates $\boldsymbol{\alpha}, \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s$ from the complete-data likelihood $p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s)$, which reduces to estimating $\boldsymbol{\beta}_i^s$ separately for each unit, if no common coefficient $\boldsymbol{\alpha}$ is present. The fixed-effects approach leads to estimates that are more dispersed than the set of parameters one is estimating. Think, for instance, of the extreme case that all $\boldsymbol{\beta}_i^s$s are actually equal. Nevertheless the individual ML estimators of $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s$ will be dispersed, with the dispersion disappearing only for T_i going to infinity.

Thus even for a likelihood-based approach it has been long recommended to consider the so-called random-effects approach where it is assumed that $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s$ are drawn independently from an underlying distribution $p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta})$, which may depend on some hyperparameter $\boldsymbol{\vartheta}$, therefore:

$$p(\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s|\boldsymbol{\vartheta}) = \prod_{i=1}^N p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta}).$$

By combining model (8.43) with one the distributions $p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta})$ discussed earlier in Subsection 8.5.1 different useful models emerge. An early reference that shows how pooling helps in problems of simultaneous inference on a set of related parameters $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s$ is Rao (1975); see also Efron and Morris (1977) for some enlightening discussion.

In a Bayesian approach, the distribution $p(\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s|\boldsymbol{\vartheta})$ takes the role of a prior distribution which is combined with observations arising from model (8.43) through Bayes' theorem; see Lindley and Smith (1972).

The Hierarchical Bayes Model

The standard mixed-effects model introduced in Laird and Ware (1982), and applied in many subsequent papers, results from combining model (8.43) with the normal distribution of heterogeneity

$$\boldsymbol{\beta}_i^s \sim \mathcal{N}_d\left(\boldsymbol{\beta}, \mathbf{Q}\right), \tag{8.45}$$

where $\boldsymbol{\beta}$ and $\mathbf{Q}$ are unknown parameters. Morris (1983) discusses that such a prior allows borrowing strength from the ensemble, when estimating $\boldsymbol{\beta}_i^s$ which is shrunken toward the population mean $\boldsymbol{\beta}$. In marketing research this model is also known as the hierarchical Bayes model; see, for instance, Rossi et al. (2005, Chapter 5). If we rewrite (8.45) as $\boldsymbol{\beta}_i^s = \boldsymbol{\beta} + \mathbf{w}_i$, $\mathbf{w}_i \sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{Q}\right)$, and substitute into (8.43), we obtain:

$$\mathbf{y}_i = \mathbf{X}_i^f\boldsymbol{\alpha} + \mathbf{X}_i^r\boldsymbol{\beta} + \mathbf{X}_i^r\mathbf{w}_i + \boldsymbol{\varepsilon}_i.$$

Under the common assumption that $\mathbf{w}_i$ and $\boldsymbol{\varepsilon}_i$ are independent, the hierarchical Bayes model corresponds to the following multivariate regression model,

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}_i, \qquad \tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \mathbf{V}_i\right), \tag{8.46}$$

with constrained error variance–covariance matrix

$$\mathbf{V}_i = \mathbf{X}_i^r \mathbf{Q} (\mathbf{X}_i^r)' + \sigma_\varepsilon^2 / \omega_i \mathbf{I}_{T_i}.$$

Subsequently, model (8.46) is referred to as the *marginal model*, because the random coefficients $\boldsymbol{\beta}_i^s$ no longer appear in this specification. The marginal model clearly indicates that despite allowing for heterogeneity the hierarchical Bayes model implies the rather inflexible normal distribution as a marginal distribution for $\mathbf{y}_i$. Further issues, in particular estimation of this widely used model, are well discussed in the many excellent monographs mentioned at the beginning of this section.

Verbeke and Lesaffre (1997) study the effect of misspecifying the random effect distribution in the linear mixed-effects model. They show that the normal shrinkage prior (8.45) yields consistent estimates of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Q}$, and σ_ε^2 even if the random effects are not normal, however, standard errors need to be corrected.

The Latent Class Regression Model

More flexibility in the marginal distribution of $\mathbf{y}_i$ is achieved by assuming that the distribution $p(\boldsymbol{\beta}_i^s|\boldsymbol{\vartheta})$ is a discrete distribution with K unknown support points $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ with $\Pr(\boldsymbol{\beta}_i^s = \boldsymbol{\beta}_k) = \eta_k$. In this case, the marginal distribution of $\mathbf{y}_i$ is the following finite mixture distribution,

$$p(\mathbf{y}_i|\omega_i, \boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k f_N(\mathbf{y}_i; \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_k, \sigma_\varepsilon^2/\omega_i \mathbf{I}_{T_i}).$$

By introducing the hidden allocation variable S_i, which takes the value k, iff $\boldsymbol{\beta}_i^s = \boldsymbol{\beta}_k$, the model may be written as the following finite mixture of multivariate mixed-effects regression models,

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_{S_i} + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \sigma_\varepsilon^2/\omega_i \mathbf{I}_{T_i}\right), \tag{8.47}$$

which is an extension of the finite mixture regression model discussed in Section 8.4 to multivariate observations $\mathbf{y}_i$. This model is also called the latent class regression model, as conditional on knowing S_i and ω_i the observations $y_{i1}, \ldots, y_{i,T_t}$ are independent.

Many interesting applications of this model are found in marketing research; see, for instance, DeSarbo et al. (1992) for metric conjoint analysis, Ramaswamy et al. (1993) for latent pooling of marketing mix elasticities, as well as Wedel and Steenkamp (1991) and the review in Wedel and DeSarbo (1993b).

The Heterogeneity Model

A very general model results if the observation model (8.43) is combined with a heterogeneity distribution assumed to be a mixture of multivariate normal distributions:

$$\boldsymbol{\beta}_i^s \sim \sum_{k=1}^{K} \eta_k \mathcal{N}_d \left(\boldsymbol{\beta}_k, \mathbf{Q}_k\right), \tag{8.48}$$

with unknown component means $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, unknown component variance–covariance matrices $\mathbf{Q}_1, \ldots,$ $\mathbf{Q}_K$, and unknown weight distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$. A constrained version of this model with $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$ being the same for all components was introduced by Verbeke and Lesaffre (1996) for homogeneous error variances. A similar model is discussed in Allenby et al. (1998), however, without considering fixed effects. Lenk and DeSarbo (2000) extend this model to observations from distributions from general exponential families; see also Section 9.6.2. Verbeke and Molenberghs (2000) introduced the terminology *heterogeneity model* for this model.

The heterogeneity model encompasses the other models discussed above. If $\mathbf{Q}_k$ is equal to a null matrix in all groups, the latent class regression model results, whereas the hierarchical Bayes model results as that special case where $K = 1$. After introducing the allocation variable S_i in the finite mixture distribution (8.48), the following distribution of heterogeneity results conditional on holding S_i fixed,

$$\boldsymbol{\beta}_i^s | S_i \sim \mathcal{N}_d \left(\boldsymbol{\beta}_{S_i}, \mathbf{Q}_{S_i}\right).$$

Because the N units form K groups, where within each group heterogeneity is described by a group-specific normal distribution, the heterogeneity model may be regarded as a mixture of random-effects models.

The marginal model where the random effects are integrated out, while still conditioning on S_i and ω_i, reads:

$$\mathbf{y}_i = \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_{S_i} + \tilde{\boldsymbol{\varepsilon}}_i, \qquad \tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_{T_i} \left(\mathbf{0}, \mathbf{V}_i\right), \tag{8.49}$$

where

$$\mathbf{V}_i = \mathbf{X}_i^r \mathbf{Q}_{S_i} (\mathbf{X}_i^r)' + \sigma_\varepsilon^2 / \omega_i \mathbf{I}_{T_i}. \tag{8.50}$$

Therefore the heterogeneity model may also be regarded as a finite mixture of multivariate mixed-effects regression models, where the errors within each unit are correlated, as opposed to the latent class regression model, where these errors are uncorrelated.

The model found applications in marketing to deal with preference heterogeneity of consumers (Allenby et al., 1998; Otter et al., 2004), in economics to analyze individual records of work and life history data (Oskrochi and Davies, 1997) and to find convergence clubs in a macroeconomic panel (Canova, 2004;

Frühwirth-Schnatter and Kaufmann, 2006b), and in biology to analyze microarray data (Lopes et al., 2003). An extension of this model which includes a dynamic linear trend model is studied in Gamerman and Smith (1996). Nobile and Green (2000) apply a modification of this model with separate random effects, each following a mixture of normal distributions, to estimate main and interaction effects in a factorial experiment.

8.5.3 Choosing the Prior for Bayesian Estimation

For Bayesian estimation, a prior on $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \boldsymbol{\eta}, \boldsymbol{\alpha}, \sigma_\varepsilon^2)$ has to be chosen. Because $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \boldsymbol{\eta})$ are unknown parameters in a finite mixture of multivariate normal distributions, the same priors as in Subsection 6.3.2 may be applied.

One could choose a conditionally conjugate prior for $\boldsymbol{\beta}_k$ where the prior variance depends on $\mathbf{Q}_k$, $\mathbf{B}_{0,k} = \mathbf{Q}_k/N_0$. On the other hand, in the marginal model (8.49), where the random effects are integrated out, $\boldsymbol{\beta}_k$ appears as a regression coefficient in a finite mixture of regression models, where no conditionally conjugate prior variance exists due to the correlation in the errors. This suggests choosing $\mathbf{B}_0$ independent of $\mathbf{Q}_k$.

$\boldsymbol{\alpha}$ and σ_ε^2 have a similar meaning as for a finite mixture of mixed-effects regression models, therefore the prior is chosen as in Subsection 8.4.2. The joint prior reads:

$$\begin{aligned} &\boldsymbol{\beta}_k \sim \mathcal{N}_d(\mathbf{b}_0, \mathbf{B}_0), \qquad && \mathbf{Q}_k^{-1} \sim \mathcal{W}_d\left(c_0^Q, \mathbf{C}_0^Q\right), \\ &\boldsymbol{\alpha} \sim \mathcal{N}_r(\mathbf{a}_0, \mathbf{A}_0), && \sigma_\varepsilon^2 \sim \mathcal{G}^{-1}(c_0^\varepsilon, C_0^\varepsilon), \\ &\boldsymbol{\eta} \sim \mathcal{D}(e_0, \ldots, e_0). \end{aligned} \tag{8.51}$$

8.5.4 Bayesian Parameter Estimation When the Allocations Are Known

For a general Bayesian analysis of the heterogeneity model it is helpful to start with parameter estimation, when the allocations $\mathbf{S} = (S_1, \ldots, S_N)$ as well as the variance parameters $\mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2$ and $\boldsymbol{\omega}$ are known.

Then the joint posterior of the regression parameters $\boldsymbol{\alpha}^\star = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ and the random coefficients $\boldsymbol{\beta}^s = (\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s)$ partitions as follows,

$$\begin{aligned} &p(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^s | \mathbf{y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2, \boldsymbol{\omega}, \mathbf{S}) \\ &= p(\boldsymbol{\alpha}^\star | \mathbf{y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2, \boldsymbol{\omega}, \mathbf{S}) \prod_{i=1}^N p(\boldsymbol{\beta}_i^s | \mathbf{y}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}_{S_i}, \mathbf{Q}_{S_i}, \omega_i) . \end{aligned}$$

Conditional on knowing the fixed effects, the random coefficients $\boldsymbol{\beta}_i^s$ are independent. Because the allocations $\mathbf{S}$ are known, the prior of $\boldsymbol{\beta}_i^s$ is normal,

$$\boldsymbol{\beta}_i^s \sim \mathcal{N}_d\left(\boldsymbol{\beta}_{S_i}, \mathbf{Q}_{S_i}\right),$$

whereas the complete-data likelihood results from:

$$\mathbf{y}_i - \mathbf{X}_i^f \boldsymbol{\alpha} = \mathbf{X}_i^r \boldsymbol{\beta}_i^s + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \sigma_\varepsilon^2/\omega_i \mathbf{I}_{T_i}\right).$$

Combining these two sources of information yields the following posterior of $\boldsymbol{\beta}_i^s$,

$$\boldsymbol{\beta}_i^s | \mathbf{y}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}_{S_i}, \mathbf{Q}_{S_i}, \omega_i \sim \mathcal{N}_d\left(\mathbf{b}_i^s, \mathbf{B}_i^s\right),$$

where the moments are given in terms of an information filter:

$$\begin{aligned} \mathbf{B}_i^s &= (\mathbf{Q}_{S_i}^{-1} + (\mathbf{X}_i^r)' \mathbf{X}_i^r \omega_i/\sigma_\varepsilon^2)^{-1}, \\ \mathbf{b}_i^s &= \mathbf{B}_i^s(\mathbf{Q}_{S_i}^{-1}\boldsymbol{\beta}_{S_i} + (\mathbf{X}_i^r)'(\mathbf{y}_i - \mathbf{X}_i^f\boldsymbol{\alpha})\omega_i/\sigma_\varepsilon^2). \end{aligned} \tag{8.52}$$

If $T_i < d$, it is more efficient to work with the following filter form which is derived in Subsection 13.3.2,

$$\begin{aligned} \mathbf{b}_i^s &= \boldsymbol{\beta}_{S_i} + \mathbf{K}_i(\mathbf{y}_i - \mathbf{X}_i^f\boldsymbol{\alpha} - \mathbf{X}_i^r\boldsymbol{\beta}_{S_i}), \\ \mathbf{B}_i^s &= (\mathbf{I}_{T_i} - \mathbf{K}_i\mathbf{X}_i^r)\mathbf{Q}_{S_i}, \\ \mathbf{K}_i &= \mathbf{Q}_{S_i}(\mathbf{X}_i^r)'\mathbf{V}_i^{-1}, \end{aligned} \tag{8.53}$$

with $\mathbf{V}_i$ being the error variance–covariance matrix of the marginal model defined in (8.50).

The posterior $p(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2, \boldsymbol{\omega}, \mathbf{S})$ is a conditional distribution, where the allocations are known, whereas the random coefficients $\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s$ are unknown. The prior of $\boldsymbol{\alpha}^*$ is a normal distribution, $\boldsymbol{\alpha}^* \sim \mathcal{N}_{r^*}(\mathbf{a}_0^*, \mathbf{A}_0^*)$, where $r^* = r + Kd$ and $\mathbf{a}_0^*$ and $\mathbf{A}_0^*$ are derived in an obvious way from the parameters $\mathbf{a}_0, \mathbf{A}_0$, $\mathbf{b}_0$, and $\mathbf{B}_0$ of the prior defined in (8.51). This prior is combined with the likelihood function $p(\mathbf{y}|\boldsymbol{\alpha}^*, \mathbf{y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2, \boldsymbol{\omega}, \mathbf{S})$ of the marginal model (8.49), where the random effects are integrated out.

The posterior distribution $p(\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2, \boldsymbol{\omega}, \mathbf{S})$ is derived in a similar way as in Subsection 8.4.3, which concerned finite mixtures of multiple mixed-effects models, whereas in the present case we are dealing with a multivariate one. By introducing a dummy coding for S_i through K binary variables $D_{ik}, k = 1, \ldots, K$, where $D_{ik} = 1$, iff $S_i = k$, and 0 otherwise, we rewrite the marginal model (8.49) as

$$\mathbf{y}_i = \boldsymbol{Z}_i^* \boldsymbol{\alpha}^* + \tilde{\boldsymbol{\varepsilon}}_i, \qquad \tilde{\boldsymbol{\varepsilon}}_i \sim \mathcal{N}_{T_i}\left(\mathbf{0}, \mathbf{V}_i\right), \tag{8.54}$$

where the design matrix $\boldsymbol{Z}_i^*$ is defined as

$$\boldsymbol{Z}_i^* = \left(\mathbf{X}_i^f \ \mathbf{X}_i^r D_{i1} \ldots \mathbf{X}_i^r D_{iK}\right).$$

Because model (8.54) is a multivariate regression model, the posterior of $\boldsymbol{\alpha}^*$ arises from a normal distribution:

$$\boldsymbol{\alpha}^*|\mathbf{y}, \mathbf{Q}_1, \ldots, \mathbf{Q}_K, \sigma_\varepsilon^2, \boldsymbol{\omega}, \mathbf{S} \sim \mathcal{N}_{r^*}\left(\mathbf{a}_N^*, \mathbf{A}_N^*\right),$$

where

$$(\mathbf{A}_N^*)^{-1} = \sum_{i=1}^{N} (\boldsymbol{Z}_i^*)' \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i^* + (\mathbf{A}_0^*)^{-1},$$

$$\mathbf{a}_N^* = (\mathbf{A}_N^*)^{-1} \left(\sum_{i=1}^{N} (\boldsymbol{Z}_i^*)' \boldsymbol{V}_i^{-1} \mathbf{y}_i + (\mathbf{A}_0^*)^{-1} \mathbf{a}_0^* \right).$$

8.5.5 Practical Bayesian Estimation Using MCMC

Empirical Bayesian estimation of the heterogeneity model, including classification, is discussed in Verbeke and Lesaffre (1996). A fully Bayesian analysis of the heterogeneity model for a fixed number K of groups via MCMC methods is discussed by Allenby et al. (1998), Lenk and DeSarbo (2000), and Frühwirth-Schnatter et al. (2004).

Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ denote all observations. MCMC estimation of the most general model is based on three levels of data augmentation. First, one introduces the discrete latent group indicators $\mathbf{S} = (S_1, \ldots, S_N)$, with S_i taking values in $\{1, \ldots, K\}$ and thereby indicating to which group unit i belongs. Second, the vector of unknowns is augmented by the random effects $\boldsymbol{\beta}^s = (\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_N^s)$. And finally, under heterogeneous error variances the scale factors $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_N)$ are added in a third data augmentation step. The joint posterior distribution of all unknowns reads:

$$p(\boldsymbol{\vartheta}, \boldsymbol{\beta}^s, \mathbf{S}, \boldsymbol{\omega}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\omega}, \boldsymbol{\beta}^s, \boldsymbol{\alpha}, \sigma_\varepsilon^2) p(\boldsymbol{\omega}) p(\boldsymbol{\alpha}, \sigma_\varepsilon^2)$$
$$\times p(\boldsymbol{\beta}^s|\mathbf{S}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{Q}_1, \ldots, \mathbf{Q}_K) p(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{Q}_1, \ldots, \mathbf{Q}_K) p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}).$$

A straightforward way of Bayesian estimation of the heterogeneity model via MCMC methods is Gibbs sampling from full conditional distributions. The sampler is discussed in Allenby et al. (1998) and Lenk and DeSarbo (2000) for a heterogeneity model with homogeneous error variances and draws in turn $\boldsymbol{\alpha}$, σ_ε^2, $\boldsymbol{\eta}$, $\boldsymbol{\beta}_k$, and $\mathbf{Q}_k$ for $k = 1, \ldots, K$, and S_i and $\boldsymbol{\beta}_i^s$ for $i = 1, \ldots, N$, from the appropriate full conditional distributions given the remaining parameters and the data $\mathbf{y}$.

It has been demonstrated in Frühwirth-Schnatter et al. (2004) that the full conditional Gibbs sampler is sensitive to the way model (8.43) is parameterized, depending on whether $\mathbf{X}_i^f$ and $\mathbf{X}_i^r$ have common columns. Sensitivity of Gibbs sampling with respect to parameterizing the standard mixed-effects model was noted earlier by Gelfand et al. (1995), and several papers show that marginalization helps in improving the performance of the Gibbs sampler; see, for instance, Meng and Van Dyk (1997, 1999), Chib and Carlin (1999), and van Dyk and Meng (2001).

The partly marginalized Gibbs sampler suggested in Frühwirth-Schnatter et al. (2004) for homogeneous error variances draws $\mathbf{S}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ from conditional distributions where the random effects $\boldsymbol{\beta}^s$ are integrated out. This sampler is shown to be less sensitive to the parameterization and has been extended in Frühwirth-Schnatter et al. (2005) to deal with heterogeneous error variances. It is summarized in the following algorithm.

Algorithm 8.3: MCMC Estimation of the Heterogeneity Model

(a) Parameter simulation conditional on the allocations $\mathbf{S}$, the random effects $\boldsymbol{\beta}^s$ and the scaling factors $\boldsymbol{\omega}$.
  (a1) Sample $\boldsymbol{\eta}$ from the conditional Dirichlet posterior $p(\boldsymbol{\eta}|\mathbf{S})$ as in *Algorithm 3.4*.
  (a2) Sample all regression coefficients $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ jointly from the posterior distribution $\boldsymbol{\alpha}^* \sim \mathcal{N}_{r^\star}(\mathbf{a}_N^*, \mathbf{A}_N^*)$, derived conditional on $\mathbf{y}, \mathbf{S}$, $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$, σ_ε^2, and $\boldsymbol{\omega}$.
  (a3) Sample each variance–covariance matrix $\mathbf{Q}_k$, $k = 1, \ldots, K$, from the posterior distribution $\mathbf{Q}_k^{-1} \sim \mathcal{W}_d\left(c_k^Q, \mathbf{C}_k^Q\right)$, derived conditional on $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\beta}^s$, and $\mathbf{S}$.
  (a4) Sample σ_ε^2 from $\mathcal{G}^{-1}(c_N^\varepsilon, C_N^\varepsilon)$, derived conditional on $\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}^s$, and $\boldsymbol{\omega}$.
(b) Classification of each unit based on $\mathbf{y}_i$, $\boldsymbol{\omega}$, and $\boldsymbol{\vartheta}$: sample each element S_i of $\mathbf{S}$ from the conditional posterior $p(S_i|\boldsymbol{\vartheta}, \mathbf{y}_i, \omega_i)$ given by

$$\Pr(S_i = k|\boldsymbol{\vartheta}, \mathbf{y}_i, \omega_i) \propto \eta_k f_N(\mathbf{y}_i; \mathbf{X}_i^f \boldsymbol{\alpha} + \mathbf{X}_i^r \boldsymbol{\beta}_k, \mathbf{V}_i), \tag{8.55}$$

  where $\mathbf{V}_i$ has been defined in (8.50).
(c) Dealing with parameter heterogeneity: sample each random coefficient $\boldsymbol{\beta}_i^s$ for $i = 1, \ldots, N$ from the $\mathcal{N}_d(\mathbf{b}_i^s, \mathbf{B}_i^s)$-distribution, derived conditional on $\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \mathbf{S}$, $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$, σ_ε^2, and $\boldsymbol{\omega}$.
(d) Dealing with variance heterogeneity: sample each scaling factor ω_i from the $\mathcal{G}(c_i^\omega, C_i^\omega)$ distribution, derived conditional on $\mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}^s$, and σ_ε^2.

Estimation of the regression coefficients $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ in step (a2) is based on the marginal model (8.49) where the random effects are integrated out in order to improve the mixing properties of the sampler. The appropriate moments were derived in Subsection 8.5.4.

Sampling of the covariance matrices $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$ in step (a3) follows immediately from *Algorithm 6.2*, dealing with mixtures of normal distributions, because the random effects $\boldsymbol{\beta}_i^s$ are assumed to be known in this step. The precise form of c_k^Q and $\mathbf{C}_k^Q$ depends upon the chosen prior covariance matrix $\mathbf{B}_0$. If $\mathbf{B}_0$ is independent of $\mathbf{Q}_k$, then

$$c_k^Q = c_0^Q + \frac{N_k}{2},$$
$$\mathbf{C}_k^Q = \mathbf{C}_0^Q + \frac{1}{2} \sum_{i:S_i=k} (\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_k)(\boldsymbol{\beta}_i^s - \boldsymbol{\beta}_k)',$$

where $N_k = \#\{S_i = k\}$.

The appropriate posterior distribution in step (a4) is easily derived from the complete-data likelihood function, which reads:

$$p(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\beta}^s,\boldsymbol{\omega}) = \prod_{i=1}^{N}\left(\frac{\omega_i}{2\pi\sigma^2_\varepsilon}\right)^{T_i/2} \times \exp\left(-\frac{1}{2\sigma^2_\varepsilon}\sum_{i=1}^{N}\omega_i\|\mathbf{y}_i-\mathbf{X}^f_i\boldsymbol{\alpha}-\mathbf{X}^r_i\boldsymbol{\beta}^s_i\|^2_2\right). \tag{8.56}$$

Therefore:

$$c^\varepsilon_N = c^\varepsilon_0 + \frac{1}{2}\left(\sum_{i=1}^{N}T_i\right),$$
$$C^\varepsilon_N = C^\varepsilon_0 + \frac{1}{2}\left(\sum_{i=1}^{N}\omega_i\|\mathbf{y}_i-\mathbf{X}^f_i\boldsymbol{\alpha}-\mathbf{X}^r_i\boldsymbol{\beta}^s_i\|^2_2\right).$$

In step (b), the indicators $S_1,\dots,S_N$ are conditionally independent given $\mathbf{y}$, $\boldsymbol{\omega}$, and $\boldsymbol{\vartheta}$, as it is assumed that the units are drawn randomly from the underlying population. The classification rule (8.55) is based on the marginal model (8.49), where the random effects are integrated out, in order to improve the mixing properties of the sampler.

In step (c), the moments of the $\mathcal{N}_d\left(\mathbf{b}^s_i,\mathbf{B}^s_i\right)$ distribution to sample the random effects are given by (8.52) or (8.53).

Finally, the posterior in step (d) follows immediately from the complete-data likelihood given in (8.56) in combination with the prior (8.42):

$$c^\omega_i = \frac{\nu}{2}+\frac{T_i}{2}, \qquad C^\omega_i = \frac{\nu}{2}+\frac{1}{2\sigma^2_\varepsilon}\|\mathbf{y}_i-\mathbf{X}^f_i\boldsymbol{\alpha}-\mathbf{X}^r_i\boldsymbol{\beta}^s_i\|^2_2.$$

8.5.6 Dealing with Model Specification Uncertainty

BIC or Schwarz criterion is quite popular for model selection in random-effect models, however, problems are reported in Stone (1974) and McCulloch and Rossi (1992) for few repeated measurements with large heterogeneity, where the number of parameters actually grows with N.

Watier et al. (1999) and Nobile and Green (2000) extend the reversible jump MCMC method of Richardson and Green (1997) to select the unknown number of components in a finite mixture of random-effects models.

Marginal likelihoods for selecting between the different models were considered by Lenk and DeSarbo (2000) and Frühwirth-Schnatter et al. (2004, 2005). Marginal likelihoods allow not only choosing the number of components, but also a comparison between the different types of heterogeneity distributions; see also the case study in Subsection 8.5.7.

Table 8.1. Marketing Data, logarithm of marginal likelihoods $p(\mathbf{y}|K,\nu)$ (from Frühwirth-Schnatter et al. (2005) with permission granted by Springer-Verlag, Wien)

	$\log p(\mathbf{y}\|K,\nu=\infty)$		$\log p(\mathbf{y}\|K,\nu=4)$	
K	$\mathbf{Q}_k \neq \mathbf{O}$	$\mathbf{Q}_k = \mathbf{O}$	$\mathbf{Q}_k \neq \mathbf{O}$	$\mathbf{Q}_k = \mathbf{O}$
1	–9222.36	–10077.31	–9101.52	–9980.21
2	–9165.66	–9881.49	**–9028.81**	–9727.13
3	**–9161.27**	–9733.98	–9043.96	–9576.97
4	–9165.73	–9669.98	–9045.86	–9522.18
5	—	–9596.61	—	–9453.22
⋮				
12	—	—	—	–9332.96
13	—	—	—	–9329.49
14	—	—	—	**–9326.26**
15	—	—	—	–9327.27
16	—	–9464.77	—	—
17	—	**–9460.61**	—	—
18	—	–9465.79	—	—

8.5.7 Application to the Marketing Data

This application concerns conjoint analysis in marketing, a procedure that is focused on obtaining the importance of certain product attributes and their significance in motivating a consumer toward purchase from a holistic appraisal of attribute combinations.

The Marketing Data come from a brand–price trade-off study in the mineral water market. Each of 213 Austrian consumers evaluated their likelihood of purchasing 15 different product-profiles offering five different brands of mineral water at different prices on 20-point rating scales. The goal of the modeling exercise is to find a model describing consumers' heterogeneous preferences toward the different brands of mineral water and their brand–price trade-offs. These data were analyzed in several studies based on homogeneous errors using a random coefficient model (Frühwirth-Schnatter and Otter, 1999), the latent class regression model (Otter et al., 2002), and the heterogeneity model (Otter et al., 2004). The material in this subsection is based on Frühwirth-Schnatter et al. (2005), where these models are compared to models based on unit-specific variance heterogeneity.

The design matrix consists of 15 columns corresponding to the constant, the four brands Römerquelle (*RQ*), Vöslauer (*VOE*), Juvina (*JU*), and Waldquelle (*WA*), a linear and a quadratic price effect, and four brand by linear price and four brand by quadratic price interaction effects. A dummy coding is used for the brands, hence the fifth brand Kronsteiner (*KR*) was chosen as the baseline. The smallest price is subtracted from the linear price column; the quadratic price is a contrast from the centered linear price. Therefore, the constant corresponds to the purchase likelihood of Kronsteiner at the lowest price level, if quadratic price effects are not present. The investigations of these data in Otter et al. (2002) indicated that a specification with fixed brand by quadratic price interactions is preferable, therefore the dimension of $\boldsymbol{\beta}_k$ is equal to $d = 11$, whereas the dimension of $\boldsymbol{\alpha}$ is equal to $r = 4$.

The prior is chosen as in Subsection 8.5.3. $\mathbf{a}_0$ and $\mathbf{b}_0$ are equal to the population mean of the random coefficient model reported in Frühwirth-Schnatter and Otter (1999), whereas $\mathbf{A}_0^{-1} = 0.04 \times \mathbf{I}_4$ and $\mathbf{B}_0^{-1} = 0.04 \times \mathbf{I}_{11}$. In the prior of $\mathbf{Q}_k$, $c_0^Q = 10$ whereas $\mathbf{C}_0^Q$ is derived by matching the prior mean, $E(\mathbf{Q}_k) = (c_0^Q - (d+1)/2)^{-1}\mathbf{C}_0^Q$, to a sample estimate computed from individual OLS estimation. In the prior of σ_ε^2, $c_0^\varepsilon = C_0^\varepsilon = 0$, whereas the prior on $\boldsymbol{\eta}$ is a $\mathcal{D}(1, \dots, 1)$ distribution.

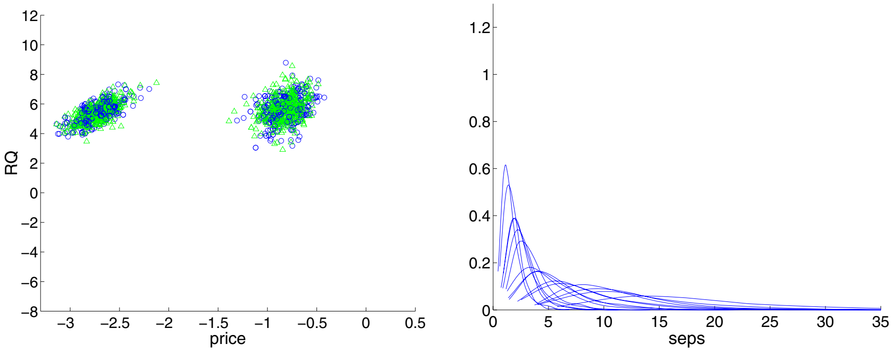

Fig. 8.1. Marketing Data; heterogeneity model with $K = 2$ and heterogeneous variances with $\nu = 4$, scatter plot of price against brand RQ (left-hand side) and posterior distribution of individual variances $\sigma_\varepsilon^2/\omega_i$ for 15 randomly selected consumers (from Frühwirth-Schnatter et al. (2005) with permission granted by Springer-Verlag, Wien)

The following finite mixture models with $K > 1$ were fitted to these data with varying the number K of groups: the general heterogeneity model, where $\mathbf{Q}_k \neq \mathbf{O}$ for all $k = 1, \dots, K$ and the latent class regression model, where $\mathbf{Q}_k = \mathbf{O}$ for all $k = 1, \dots, K$. These models were compared to the hierarchical Bayes model, which formally corresponds to a heterogeneity model with $K = 1$ and $\mathbf{Q}_1 \neq \mathbf{O}$. Each of these models was fitted with heterogeneous variances with $\nu = 4$ as well as with homogeneous variances that correspond to $\nu = \infty$. Estimation was carried through 30,000 MCMC iterations, with the last 6000 draws being kept for inference.

Table 8.1 shows estimates of the logarithm of the marginal likelihood $p(\mathbf{y}|K, \nu)$ for various models obtained by bridge sampling. The hierarchical Bayes model (column $\mathbf{Q}_k \neq \mathbf{O}$, line $K = 1$) is clearly preferred to all latent class regression models (column $\mathbf{Q}_k = \mathbf{O}$), but is outperformed by the heterogeneity model (column $\mathbf{Q}_k \neq \mathbf{O}$, lines with $K > 1$), regardless of the assumption made concerning the variances.

The specification chosen for the variance exercises a considerable influence on the number of optimal classes. Under the assumption of homogeneous variances the optimal latent class regression model has seventeen classes, whereas the number reduces to fourteen under heterogeneous errors. Also the heterogeneity model has a different number of optimal classes, namely two under heterogeneous errors and three under homogeneous errors. The optimal model of all models under consideration is a heterogeneity model with heterogeneous error variances and $K = 2$ classes. The preference of a model with heterogeneous variances is also supported by Figure 8.1, which shows considerable differences in the posterior distribution of the individual variances $\sigma_\varepsilon^2/\omega_i$ for 15 randomly selected consumers.

Table 8.2. Marketing Data, heterogeneity model with $K = 2$ and heterogeneous variances with $\nu = 4$; posterior expectation of the group-specific parameters $\boldsymbol{\beta}_k$ and the group-specific weights η_k; posterior standard deviations in parentheses (from Frühwirth-Schnatter et al. (2005) with permission granted by Springer-Verlag, Wien)

	$\beta_{k,j}$			$\beta_{k,j}$	
	$k = 1$	$k = 2$		$k = 1$	$k = 2$
const	14.78	12.43	$RQ \times p$	–0.71	–0.04
	(0.67)	(0.75)		(0.16)	(0.15)
RQ	5.44	5.65	$VOE \times p$	–0.85	–0.02
	(0.65)	(0.84)		(0.16)	(0.16)
VOE	5.30	5.17	$JU \times p$	–0.38	0.07
	(0.65)	(0.97)		(0.16)	(0.16)
JU	1.28	0.38	$WA \times p$	–0.58	–0.10
	(0.66)	(0.97)		(0.15)	(0.13)
WA	2.24	1.10			
	(0.68)	(0.78)			
p	–2.72	–0.82		η_k	
	(0.15)	(0.15)		$k = 1$	$k = 2$
p^2	–0.03	0		0.58	0.42
	(0.07)	(0.06)		(0.04)	(0.04)

We proceed with estimating the group-specific parameters for this model. The posterior draws in Figure 8.1 are the point process representation of the projection onto the coefficients $\beta_{k,2}$ and $\beta_{k,6}$ which correspond to the effect of the brand *RQ* and the price effect. We find two clearly separated simulation clusters, with one group collecting very price-sensitive consumers whereas the consumers of the other group are less price sensitive. Therefore it is possible to identify the model through putting the constraint $\beta_{1,6} < \beta_{2,6}$

on the group-specific price coefficient. Table 8.2 gives the resulting estimates for the group-specific parameters and the group weights.

8.6 Further Issues

8.6.1 Regression Modeling Based on Multivariate Mixtures of Normals

Müller et al. (1996) show that for stochastic regressor variables finite mixtures of multivariate normal distributions could be used as an alternative tool for flexible regression modeling. Consider, for example, a bivariate random variable (X, Y), modeled by a mixture of bivariate normal distributions with component means $\boldsymbol{\mu}_k$, component covariance matrices $\boldsymbol{\Sigma}_k$, and weight distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$.

The conditional density $p(y|X = x_i, \boldsymbol{\vartheta})$ of Y given $X = x_i$ is easily found to be equal to the following univariate mixture of normal distributions,

$$p(y|X = x_i, \boldsymbol{\vartheta}) = \tag{8.57}$$
$$\sum_{k=1}^{K} w_k(x_i, \boldsymbol{\vartheta}) f_N(y; \beta_{k,1} x_i + \beta_{k,2}, \Sigma_{k,22}(1 - \rho_k^2)),$$

where

$$\beta_{k,1} = \rho_k \sqrt{\frac{\Sigma_{k,11}}{\Sigma_{k,22}}}, \qquad \beta_{k,2} = \mu_{k,2} - \beta_{k,1}\mu_{k,1},$$

with ρ_k being the group-specific correlation coefficient

$$\rho_k = \frac{\Sigma_{k,12}}{\sqrt{\Sigma_{k,11}\Sigma_{k,22}}},$$

and

$$w_k(x_i, \boldsymbol{\vartheta}) \propto \eta_k f_N(x_i; \mu_{k,1}, \Sigma_{k,11}).$$

Density (8.57) is closely related to the density of a finite mixture of regression models, where the slope, the intercept, and the error variance of the regression model switch among the different components. The component weights $w_k(x_i, \boldsymbol{\vartheta})$, however, are not fixed, but vary with x_i, and will be higher for components that are closer to x_i than others.

The dependence of the weights on observations is implicit in this application of a multivariate mixture distribution to a regression type analysis. Several extensions of standard finite mixture models and finite mixtures of a regression model that are based on explicitly modeling such a dependence of the weights on observations are discussed in Subsections 8.6.2 and 8.6.3.

8.6.2 Modeling the Weight Distribution

For a standard finite mixture regression model the joint distribution $p(\mathbf{y}_i, S_i|\boldsymbol{\vartheta})$ factors as $p(\mathbf{y}_i, S_i|\boldsymbol{\vartheta}) = p(\mathbf{y}_i|S_i, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)p(S_i|\boldsymbol{\eta})$, where the prior classification probabilities $\Pr(S_i = k|\boldsymbol{\eta}) = \eta_k$ are modeled as being independent of any data. In the marginal mixture distribution of $\mathbf{y}_i$ this leads to mixture density with fixed weight distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$.

Various authors suggested modeling the prior classification probabilities $\Pr(S_i = k|\boldsymbol{\eta})$ in terms of covariates $\mathbf{z}_i$; see Fair and Jaffee (1972) for an early application. This is sensible whenever the span of the covariates is different between the different clusters. A typical example is a change-point regression, where the covariate $z_i = i$ is likely to determine cluster membership.

To include covariate information, $\Pr(S_i = k|\boldsymbol{\eta})$ is first reparameterized for $k = 1, \dots, K-1$ in terms of an unconstrained parameter $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{K-1})$ using the logistic transformation:

$$\log \frac{\Pr(S_i = k|\boldsymbol{\alpha})}{\Pr(S_i = K|\boldsymbol{\alpha})} = \log \frac{\eta_k}{1 - \sum_{j=1}^{K-1} \eta_j} = \alpha_k. \qquad (8.58)$$

If for each unit i a subject-specific variable $\mathbf{z}_i$ is observed additionally to $\mathbf{y}_i$, that might help to classify the subjects, then this information could be included through a multinomial logistic regression model:

$$\log \frac{\Pr(S_i = k|\boldsymbol{\alpha}, \boldsymbol{\gamma})}{\Pr(S_i = K|\boldsymbol{\alpha}, \boldsymbol{\gamma})} = \alpha_k + \mathbf{z}_i \boldsymbol{\gamma}_k, \qquad (8.59)$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_{K-1})$ is an unknown regression parameter. If $\mathbf{z}_i$ fails to improve the resulting classification, then all components of $\boldsymbol{\gamma}$ are zero and model (8.59) reduces to (8.58). Frühwirth-Schnatter and Kaufmann (2006b) assume dependence of $\Pr(S_i = k|\boldsymbol{\alpha})$ on the initial income in an economic study involving panels of income data and use marginal likelihoods to test the more general model against a model where η_k is fixed. Scaccia and Green (2003) use time and age in a growth curve analysis to model the weight distribution in a mixture of normal distributions.

8.6.3 Mixtures-of-Experts Models

Mixtures-of-experts models have been proposed in the neural network literature by Jacobs et al. (1991), and have found widespread application for modeling relationships among variables. They are defined as the following mixture distribution,

$$p(y_i) = \sum_{k=1}^{K} \eta_{k,i} f_N(y_i; \mathbf{x}_i \boldsymbol{\beta}_k, \sigma^2_{\varepsilon,k}),$$

where

$$\text{logit } \eta_{k,i} = \alpha_k + \mathbf{x}_i \boldsymbol{\gamma}_k.$$

From a statistical point of view, such a model is a finite mixture of regression model with observation-dependent weight distribution; see again Subsection 8.6.2. Note that the mixture weights may depend on the same covariates as the mean of the regression model. This may lead to identifiability problems (Jiang and Tanner, 1999).

Jacobs et al. (1996) and Peng et al. (1996) consider Bayesian parameter estimation using MCMC. Jacobs et al. (1997) discuss Bayesian methods for model selection in mixtures-of-experts models.

Hierarchical mixtures-of-experts result if the component densities themselves are mixtures-of-experts models; see Jordan and Jacobs (1994) for estimation based on the EM algorithm and Peng et al. (1996) for a Bayesian approach.