Journal of Business & Economic Statistics

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Bayesian Analysis of the Heterogeneity Model

Sylvia Frühwirth-Schnatter, Regina Tüchler & Thomas Otter

To cite this article: Sylvia Frühwirth-Schnatter, Regina Tüchler & Thomas Otter (2004) Bayesian Analysis of the Heterogeneity Model, Journal of Business & Economic Statistics, 22:1, 2-15, DOI: 10.1198/073500103288619331

To link to this article: http://dx.doi.org/10.1198/073500103288619331

Published online: 01 Jan 2012.

Submit your article to this journal

Article views: 60

View related articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=ubes20

Download by: [KU Leuven University Library] **Date:** 13 October 2015, At: 02:28

Bayesian Analysis of the Heterogeneity Model

Sylvia Frühwirth-Schnatter
Department of Applied Statistics (IFAS), Johannes Kepler Universität, A-4040 Linz, Austria
(*Sylvia.Fruehwirth-Schnatter@jku.at*)

Regina Tüchler
Department of Statistics, University of Business Administration and Economics, A-1090 Vienna, Austria
(*regina.tuechler@wu.edu*)

Thomas Otter
Fisher College of Business, Ohio State University, 2100 Neil Avenue, Columbus, OH 43210
(*otter_2@cob.osu.edu*)

We consider Bayesian estimation of a finite mixture of models with random effects, which is also known as the heterogeneity model. First, we discuss the properties of various Markov chain Monte Carlo samplers that are obtained from full conditional Gibbs sampling by grouping and collapsing. Whereas full conditional Gibbs sampling turns out to be sensitive to the parameterization chosen for the mean structure of the model, the alternative sampler is robust in this respect. However, the logical extension of the approach to the sampling of the group variances does not further increase the efficiency of the sampler. Second, we deal with the identifiability problem due to the arbitrary labeling within the model. Finally, a case study involving metric conjoint analysis serves as a practical illustration.

KEY WORDS: Collapsing; Conjoint analysis; Grouping; Label switching; Mixture of random-effects model; Parameterization.

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

1. INTRODUCTION

In this article we consider the heterogeneity model

$$\mathbf{y}_i = \mathbf{X}_i^1 \boldsymbol{\alpha} + \mathbf{X}_i^2 \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}), \qquad i = 1, \dots, N, \tag{1}$$

where $\mathbf{y}_i$ is a vector of T_i observations for subject i, $\mathbf{X}_i^1$ is the $T_i \times d$ matrix for the $d \times 1$ vector of the fixed effects $\boldsymbol{\alpha}$, and $\mathbf{X}_i^2$ is the design matrix of dimension $T_i \times r$ for the $r \times 1$ random-effects vector $\boldsymbol{\beta}_i$. $\mathbf{I}$ is the identity matrix. $\boldsymbol{\beta}_i$ is a random effect that, due to unobserved heterogeneity, is different for each subject. The unknown distribution $\pi(\boldsymbol{\beta}_i)$ of heterogeneity is approximated by a mixture distribution,

$$\boldsymbol{\beta}_i \sim \sum_{k=1}^{K} \eta_k N(\boldsymbol{\beta}_k^G, \mathbf{Q}_k^G), \tag{2}$$

with the unknown group means $\boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_K^G$, the unknown group covariance matrices $\mathbf{Q}_1^G, \dots, \mathbf{Q}_K^G$, and the unknown group probabilities $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$. Model (1) is a finite mixture of models with random effects. This model, but with $\mathbf{Q}_1^G, \dots, \mathbf{Q}_K^G$ being the same in all groups, was introduced by Verbeke and Lesaffre (1996). Later, Verbeke and Molenberghs (2000) introduced the terminology heterogeneity model for (1). A comparable model without fixed effects has been studied by Allenby, Arora, and Ginter (1998) for logit choice models and by Lenk and DeSarbo (2000) for observations from distributions of general exponential families. In this article, however, we confine ourselves to the important special case of observations from the normal distribution.

Bayesian estimation of the heterogeneity model has been discussed by Allenby et al. (1998) and Lenk and DeSarbo (2000). Bayesian estimation starts by introducing discrete latent group indicators S_i, $i = 1, \dots, N$, taking values in $\{1, \dots, K\}$ with unknown probability distribution $\Pr(S_i = k) = \eta_k$, $k = 1, \dots, K$. With the help of the latent group indicators, model (2) is written as

$$\boldsymbol{\beta}_i \sim \begin{cases} N(\boldsymbol{\beta}_1^G, \mathbf{Q}_1^G) & \text{if } S_i = 1 \\ \vdots & \\ N(\boldsymbol{\beta}_K^G, \mathbf{Q}_K^G) & \text{if } S_i = K. \end{cases} \tag{3}$$

All unknown quantities, including the latent random effects $\boldsymbol{\beta}^N = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N)$, the latent indicators $\mathbf{S}^N = (S_1, \dots, S_N)$ and the unknown model parameters $\boldsymbol{\phi} = (\boldsymbol{\phi}_L, \boldsymbol{\phi}_V, \boldsymbol{\eta})$, are estimated jointly using Markov chain Monte Carlo (MCMC) simulation from their posterior density. We use the notation $\boldsymbol{\phi}_L = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_K^G)$ for the location parameters and $\boldsymbol{\phi}_V = (\mathbf{Q}_1^G, \dots, \mathbf{Q}_K^G, \sigma_\varepsilon^2)$ for the variance parameters. A straightforward MCMC method is to split the augmented parameter into blocks and sample the variables in each block from their full conditionals.

Algorithm 1 (Sampling from full conditionals).

(i-1) Sample $\mathbf{S}^N$ from $\pi(\mathbf{S}^N | \boldsymbol{\beta}^N, \boldsymbol{\phi}, \mathbf{y}^N)$.
(ii-1) Sample $\boldsymbol{\eta}$ from $\pi(\boldsymbol{\eta} | \mathbf{S}^N)$.
(iii-1a) Sample $\boldsymbol{\alpha}$ from $\pi(\boldsymbol{\alpha} | \boldsymbol{\beta}^N, \mathbf{S}^N, \boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_K^G, \boldsymbol{\phi}_V, \mathbf{y}^N)$.
(iii-1b) Sample $\boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_K^G$ from $\pi(\boldsymbol{\beta}_1^G, \dots, \boldsymbol{\beta}_K^G | \boldsymbol{\beta}^N, \mathbf{S}^N, \boldsymbol{\alpha}, \boldsymbol{\phi}_V, \mathbf{y}^N)$.
(iii-1c) Sample $\boldsymbol{\beta}^N$ from $\pi(\boldsymbol{\beta}^N | \mathbf{S}^N, \boldsymbol{\phi}, \mathbf{y}^N)$.
(iv-1) Sample $\boldsymbol{\phi}_V$ from $\pi(\boldsymbol{\phi}_V | \boldsymbol{\beta}^N, \mathbf{S}^N, \boldsymbol{\phi}_L, \mathbf{y}^N)$.

A comparable full conditional Gibbs sampler, but with step (ii-1) replaced by

(ii-1) Sample $\boldsymbol{\eta}$ from $\pi(\boldsymbol{\eta} | \mathbf{S}^N)$ under the restriction $\eta_1 < \cdots < \eta_K$ (constrained Algorithm 1),

© 2004 American Statistical Association
Journal of Business & Economic Statistics
January 2004, Vol. 22, No. 1
DOI 10.1198/073500103288619331

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

has been applied successfully in marketing science in the pioneering work of Allenby et al. (1998) and Lenk and DeSarbo (2000). Although Algorithm 1 happened to work fine in these applications, it is necessary to discuss potential pitfalls of this algorithm and to consider alternative methods of MCMC estimation of model (1).

A first potential drawback of Algorithm 1—either constrained or unconstrained—is that in case that correlations are high between quantities in the different subblocks, especially within step (iii), the resulting sampler will be slowly mixing. This is a well-discussed issue for the normal linear mixed model (Gelfand, Sahu, and Carlin 1995) which is a special case of the heterogeneity model with $K = 1$. We demonstrate in this article that similar problems are to be expected for the general model where $K > 1$ and the case study in Section 4.3 highlights the rather dramatic consequences for practical work. Algorithm 1 turns out to be sensitive to the parameterization used for the mean structure of model (1) and might exhibit slow convergence in cases where random effects with low variance are centered around the group means and in cases where random effects with high variance are centered around $\mathbf{0}$.

As an alternative to Algorithm 1, we exploit a common way of overcoming mixing problems caused by correlation, namely "blocking" (i.e., updating parameters jointly) and "collapsing" (i.e., sampling parameters from partially marginalized distributions) (see Chen, Shao, and Ibrahim 2000; Liu, Wong, and Kong 1994). Such blocked and collapsed samplers have been derived for a normal linear mixed model independently by Chib and Carlin (1999) and Frühwirth-Schnatter and Otter (1999); in this article, we extend them to the mixture of random-effects models (1) with $K > 1$. To our knowledge, such a collapsed sampler has not yet been considered for the general heterogeneity model. This alternative sampler will turn out to be robust to the parameterization underlying the mean structure. The resulting sampler is only partly marginalized; in the sense that all variance parameters are still sampled from their full conditional distributions. Marginalizing this part of the sampling scheme is not possible without leaving the convenient Gibbs sampling framework. However, by introducing a Metropolis–Hastings step into the sampler, we are able to investigate the merits of "collapsing" this sampling step. Interestingly, this fully marginalized MCMC scheme (in the sense that no sampling step conditions on the random effects) does not provide a noticeable improvement over the partly marginalized sampler.

The second topic of this article concerns the identifiability problem inherent in model (1). Because model (1) includes the latent, discrete structure $\mathbf{S}^N$, the unconstrained posterior typically is multimodal. The impact of this type of unidentifiability on MCMC estimation of mixture models has been discussed by Stephens (2000), Celeux, Hurn, and Robert (2000), and Frühwirth-Schnatter (2001a). In this article we add some more aspects on this identifiability problem and illustrate for simulated data that a standard constraint as the one applied in step (ii-1) of the constrained Algorithm 1 does not necessarily restrict the posterior simulations to a unique modal region. Choosing of a suitable identifiability constraint appears particularly difficult in the multivariate setting considered here. We will demonstrate within a case study from metric conjoint analysis how to explore MCMC simulations from the unconstrained posterior to obtain an identifiability constraint that is able to separate the posterior modes.

The rest of the article is organized as follows. In Section 2 we discuss, starting from Algorithm 1, various MCMC samplers obtained from Algorithm 1 by grouping and collapsing. In Section 3 we explore the unidentifiability problem. We present an empirical illustration involving metric conjoint analysis in Section 4, and conclude the article with some final remarks in Section 5.

2. PARAMETERIZATION, GROUPING, AND COLLAPSING

2.1 Parameterization and Sampling From the Full Conditionals

A straightforward method for MCMC estimation of the general heterogeneity model (1) is sampling from the full conditionals (Algorithm 1 in Sec. 1). The problem with sampling from the full conditionals is that in cases where correlations are high between quantities in different subblocks, the resulting sampler will be slowly mixing. The possible influence of the parameterization on the efficiency of the MCMC sampler, especially within step (iii-1), is a well-discussed issue for the normal linear mixed model (Gelfand et al. 1995). Because this model is the special case of (1) where $K = 1$, we expect a similar influence of the parameterization for the more general model, which, conditional on $\mathbf{S}^N$, could be regarded as the combination of K normal linear mixed models with the random effects of each group independent from each other.

2.1.1 Centered and Noncentered Parameterization. The mean structure of model (1) can be parameterized in two common ways. If $\mathbf{X}_i^1$ and $\mathbf{X}_i^2$ do not have any common column, then $\boldsymbol{\alpha}$ represents parameters that are fixed for all individuals, whereas $\boldsymbol{\beta}_i$ are the random (heterogeneous) effects centered around $\boldsymbol{\beta}_k^G$, the population mean in class k: $E(\boldsymbol{\beta}_i|S_i = k, \boldsymbol{\phi}) = \boldsymbol{\beta}_k^G$. The expectation of the marginal distribution of $\mathbf{y}_i$, where both the random effects and the latent indicators are integrated out, is given by

$$E(\mathbf{y}_i|\boldsymbol{\phi}) = \mathbf{X}_i^1\boldsymbol{\alpha} + \mathbf{X}_i^2\boldsymbol{\alpha}_{\boldsymbol{\beta}}, \qquad \boldsymbol{\alpha}_{\boldsymbol{\beta}} = \sum_{k=1}^{K} \boldsymbol{\beta}_k^G \eta_k. \tag{4}$$

Conditional on S_i, centering is around the group-specific means, whereas marginally (where the indicators are integrated out), the random effects are centered around the population mean, $E(\boldsymbol{\beta}_i|\boldsymbol{\phi}) = \boldsymbol{\alpha}_{\boldsymbol{\beta}}$. Therefore, this parameterization seems to be rather close to the idea of hierarchical centering introduced by Gelfand et al. (1995), and thus we call it *centered parameterization.* This is the parameterization used by Allenby et al. (1998) and Lenk and DeSarbo (2000).

Alternatively, Verbeke and Lesaffre (1996) designed model (1) in such a way that $\mathbf{X}_i^1$ and $\mathbf{X}_i^2$ may have common columns. This implies the additional constraint that marginally (where the indicators are integrated out) the random effects are deviations from the fixed effects with mean $\mathbf{0}$, $E(\boldsymbol{\beta}_i|\boldsymbol{\phi}) = \sum_{k=1}^{K} \boldsymbol{\beta}_k^G \eta_k = \mathbf{0}$. The first two moments of the marginal distribution of $\mathbf{y}_i$, where both the random effects and the latent

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

indicators are integrated out, are given by

$$E(\mathbf{y}_i|\boldsymbol{\phi}) = \mathbf{X}_i^1\boldsymbol{\alpha},$$

$$\operatorname{var}(\mathbf{y}_i|\boldsymbol{\phi}) = \mathbf{X}_i^2\left(\sum_{k=1}^{K} \mathbf{Q}_k^G\eta_k\right)(\mathbf{X}_i^2)' + \sigma_\varepsilon^2\mathbf{I}. \qquad (5)$$

Like in the classical random-effects model, the random effects contribute to the variance of $\mathbf{y}_i$ only. Because marginally, the random effects are centered around $\mathbf{0}$, we term this parameterization *noncentered* (as in Gelfand et al. 1995).

The two parameterizations are, of course, related. Given the centered parameterization, the noncentered parameterization is obtained by adding $\boldsymbol{\alpha}_{\boldsymbol{\beta}}$ to the fixed effects, where $\boldsymbol{\alpha}_{\boldsymbol{\beta}}$ is defined in (4) as the weighted mean of the group-specific parameters, and by subtracting $\boldsymbol{\alpha}_{\boldsymbol{\beta}}$ from each $\boldsymbol{\beta}_i$,

$$\mathbf{y}_i = \begin{bmatrix}\mathbf{X}_i^1 & \mathbf{X}_i^2\end{bmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\alpha}_{\boldsymbol{\beta}}\end{pmatrix} + \mathbf{X}_i^2(\boldsymbol{\beta}_i - \boldsymbol{\alpha}_{\boldsymbol{\beta}}) + \boldsymbol{\varepsilon}_i. \qquad (6)$$

Although the two parameterizations are theoretically equivalent, the choice of the parameterization might have a substantial influence on the efficiency of a MCMC sampler.

We discuss this in more detail for a model without fixed effects ($\mathbf{X}_i^1 = \mathbf{0}$ in the centered parameterization and $\mathbf{X}_i^1 = \mathbf{X}_i^2$ in the noncentered parameterization). In the context of a single normal linear mixed model (i.e., $K = 1$), Gelfand et al. (1995) recommended selecting the parameterization according to the contribution of the random-effects covariance matrix $\mathbf{Q}^G$ to the marginal variance of $\mathbf{y}_i$. The noncentered parameterization should be used only if the contribution of the random effects to the marginal variance is small compared with the contribution of the observation error σ_ε^2. If the variance of the random effects is dominating the variance of the marginal model, then the centered parameterization should be used.

These results are easily extended to the mixture of random-effects models considered here. Conditional on $\mathbf{S}^N$, model (1) may be regarded as a single random-effects model with heterogeneous covariance matrix $\mathbf{Q}_i = \mathbf{Q}_{S_i}^G$. The results of Gelfand et al. (1995, sec. 2), are easily extended to such a model. The crucial quantity is the determinant of the matrix

$$\mathbf{Q}_i\mathbf{B}_i^{-1} = \mathbf{Q}_i\big(\sigma_\varepsilon^{-2}(\mathbf{X}_i^2)'\mathbf{X}_i^2 + \mathbf{Q}_i^{-1}\big) = \sigma_\varepsilon^{-2}\mathbf{Q}_i(\mathbf{X}_i^2)'\mathbf{X}_i^2 + \mathbf{I}, \quad (7)$$

which reflects the contribution of the random effects to the marginal variance of $\mathbf{y}_i$. We obtain the following results. The centered parameterization is to be preferred if $|\mathbf{B}_i\mathbf{Q}_i^{-1}| = 1/|\mathbf{Q}_i\mathbf{B}_i^{-1}|$ is near 0, whereas the noncentered parameterization is to be preferred if $|\mathbf{B}_i\mathbf{Q}_i^{-1}| = 1/|\mathbf{Q}_i\mathbf{B}_i^{-1}|$ is close to 1.

These results suggest using the parameterization of Verbeke and Lesaffre (1996) in the context where $K > 1$ only in cases where the variability of the random effects is small compared to the observation error σ_ε^2 for *all groups* and to use the centered parameterization as was done by, for example, Lenk and DeSarbo (2000) in those cases where for *all groups* the variability of the random effects is large compared to the observation error σ_ε^2. Both parameterizations may be poor if for one group all effects are nearly homogeneous, whereas for the other group heterogeneity dominates. Our experiences with simulated data (reported in the next section), as well as with real data, confirm these findings.

2.1.2 Illustrative Example. Consider a simplistic metric conjoint study with a controlled full-factorial design, where N consumers evaluate two attributes of a product, one being, for instance, one of two brands, and the second being some metric attribute that is either low or high,

$$\mathbf{y}_i = \mathbf{X}_i^2\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \qquad \mathbf{X}_i^2 = \begin{pmatrix}1 & 0 & m_i^{low}\\ 1 & 1 & m_i^{low}\\ 1 & 0 & m_i^{high}\\ 1 & 1 & m_i^{high}\end{pmatrix}, \qquad (8)$$

where $\mathbf{y}_i$ is the purchase likelihood on a given scale. We assume that there are two groups of consumers with heterogeneity in the preferences,

$$\boldsymbol{\beta}_i \sim \begin{cases} N\big(\boldsymbol{\beta}_1^G, \operatorname{diag}(\delta_{11}, \delta_{12}, \delta_{13})\big) & \text{if } S_i = 1\\ N\big(\boldsymbol{\beta}_2^G, \operatorname{diag}(\delta_{21}, \delta_{22}, \delta_{23})\big) & \text{if } S_i = 2. \end{cases} \qquad (9)$$

Whether we should combine full conditional Gibbs sampling with the parameterization of Lenk and DeSarbo (2000) or rewrite the model following Lesaffre and Verbeke (1996) depends on the amount of unobserved heterogeneity. In our example,

$$(\mathbf{X}_i^2)'\mathbf{X}_i^2 = \begin{pmatrix} 4 & 2 & 2(m_i^{low} + m_i^{high})\\ 2 & 2 & m_i^{low} + m_i^{high}\\ 2(m_i^{low} + m_i^{high}) & m_i^{low} + m_i^{high} & 2(m_i^{low})^2 + 2(m_i^{high})^2 \end{pmatrix}. \qquad (10)$$

Applying the rule based on $|\mathbf{B}_i\mathbf{Q}_i^{-1}|$ [see (7)], we obtain the following result:

- Use the noncentered parameterization if for *all* groups heterogeneity is small for *all* effects compared with $\sigma_\varepsilon^2(\delta_{11}, \delta_{12}, \delta_{13}, \delta_{21}, \delta_{22}, \delta_{23} \ll \sigma_\varepsilon^2)$.
- Use the centered parameterization if for *each* group *at least one* effect dominates the marginal variance of $\mathbf{y}_i$. This result also holds if one of the other effects is nearly homogeneous (e.g., $\delta_{k3} \ll \sigma_\varepsilon^2$, whereas $\delta_{k2} \gg \sigma_\varepsilon^2$).

To illustrate these results, we simulated data and compared the output from Algorithm 1 for the two different parameterizations. The MCMC samples are compared by the lag 1 sample autocorrelation of the MCMC chain as well as by the inefficiency factor κ, which accounts for the whole serial dependence in the sampled values (Geweke 1992),

$$\kappa = 1 + 2\sum_{j=1}^{J}\left(1 - \frac{j}{J+1}\right)\rho(j), \qquad (11)$$

where $\rho(j)$ represents the autocorrelation at lag j of the sampled parameter values. κ is the factor by which we have to increase the run length of the MCMC sampler compared to iid sampling. We choose the bandwidth J such that $\rho(J)$ significantly contributes to the serial dependence of the sampled value but the succeeding $\rho(j), j = J+1, J+2, \ldots,$ are not significant. These are contained in the interval $\rho(j) \in [-2/\sqrt{M}, 2/\sqrt{M}\,]$, where M is the number of MCMC draws.

Table 1. Simulation 1, Autocorrelation at Lag 1 (AC) and Inefficiency Factors (Ineff) for the Three Sample Algorithms (Algorithms 1–3) and the Centered (C) and the Noncentered (NC) Parameterization

	Algorithm 1				*Algorithm 2*				*Algorithm 3*	
	AC		*Ineff*		*AC*		*Ineff*		*AC*	*Ineff*
	C	*NC*	*C*	*NC*	*C*	*NC*	*C*	*NC*	*C*	*C*
$\boldsymbol{\beta}_1^G$	.72	.10	19.0	4.5	.17	.11	1.8	1.4	.01	1.1
	.94	.40	34.7	10.0	.16	.14	1.6	.9	.06	1.1
	.93	.73	51.2	16.0	.04	.03	1.1	1.2	−.02	1.0
$\boldsymbol{\beta}_2^G$	.81	−.12	28.2	1.7	.08	.23	1.0	1.6	.01	1.1
	.92	.01	31.8	2.7	.04	.10	1.0	1.1	.04	1.0
	.69	.58	30.4	25.4	.14	.07	1.4	1.0	.03	1.0
$\delta_{1,1}$	.81	.85	26.6	57.1	.90	.89	5.7	14.5	.89	13.2
$\delta_{1,2}$	.86	.82	12.7	18.3	.92	.92	7.7	13.1	.91	15.7
$\delta_{1,3}$	.84	.82	9.8	9.5	.84	.89	16.2	3.8	.83	16.3
$\delta_{2,1}$	.86	.89	18.5	24.0	.92	.88	4.0	2.8	.90	13.4
$\delta_{2,2}$	.92	.89	36.0	9.0	.95	.84	41.5	14.2	.95	17.6
$\delta_{2,3}$	.82	.85	24.7	29.8	.93	.83	8.8	10.5	.90	9.4
$\boldsymbol{\eta}$	.20	−.03	13.7	2.1	.18	.20	1.5	1.9	−.02	1.0
σ_ε^2	.50	.48	17.7	13.0	.35	.38	5.1	3.0	.13	1.5

Simulation 1. The data for the first simulation come from (8) with the group specific means $\boldsymbol{\beta}_1^G = (15\ 3\ -.8)'$ and $\boldsymbol{\beta}_2^G = (7\,7\,-2)'$. We take $\mathbf{Q}_1^G = \text{diag}(1\ 1\ .5)$ and $\mathbf{Q}_2^G = \text{diag}(1.5\ 1.5\ 1)$ for the group variances and $\sigma_\varepsilon^2 = 10$ for the observation error variance. The group weights are $\eta_1 = .6$ and $\eta_2 = .4$. We simulate 200 vectors, $\mathbf{y}_i$, and run Algorithm 1 for each of the two parameterizations for 1,500 iterations. The last 1,000 iterations are kept for estimation. The simulations are based on a normal prior for the group-specific means, $\boldsymbol{\beta}_k^G \sim N((11\ 5\ -1.4)', 10\cdot\mathbf{I})$, $k = 1, 2$; an inverted Wishart prior for the group-specific covariance matrices, $\mathbf{Q}_k^G \sim IW(5, \text{diag}(3.75\ \ 3.75\ \ 3.25))$, $k = 1,2$; and a Dirichlet prior for the group weights, $\boldsymbol{\eta} \sim D(1, 1)$. We remain noninformative about the model error variance σ_ε^2. From (7) and (10), we obtain $E(|\mathbf{B}_i\mathbf{Q}_i^{-1}|) = .9198$, supporting theoretically the noncentered parameterization. This is confirmed empirically. It is obvious from the autocorrelations and the inefficiency factors of the group-specific means in Table 1 that the noncentered parameterization is more efficient than the centered parameterization. Loss of efficiency due to the centered parameterization is prominent when sampling the group-specific means $\boldsymbol{\beta}_k^G$ and $\boldsymbol{\eta}$.

Simulation 2. In this simulation we simulate data for the case where for each group the variance for at least one effect dominates the observation equations variance. We have the same group means and group weights as in simulation 1, but now take $\mathbf{Q}_1^G = \text{diag}(10\ 10\ .5)$, $\mathbf{Q}_2^G = \text{diag}(1.5\ 1.5\ 6)$, and $\sigma_\varepsilon^2 = 1$. Again we simulate 200 vectors, $\mathbf{y}_i$, and run Algorithm 1 for each of the two parameterizations for 1,500 iterations, with the final 1,000 iterations kept for estimation. We use the same prior information as for simulation 1, only the prior for the groups covariances is changed to $\mathbf{Q}_k^G \sim IW(5, \text{diag}(18\ 18\ 9))$, $k = 1, 2$. From (7) and (10), we obtain $E(|\mathbf{B}_i\mathbf{Q}_i^{-1}|) = .0012$, supporting theoretically the centered parameterization. The autocorrelations and the inefficiency factors in Table 2 clearly indicate that the centered parameterization works better.

Table 2. Simulation 2, Autocorrelation at Lag 1 (AC) and Inefficiency Factors (Ineff) for the Three Sample Algorithms (Algorithms 1–3) and the Centered (C) and the Noncentered (NC) Parameterization

	Algorithm 1				*Algorithm 2*				*Algorithm 3*	
	AC		*Ineff*		*AC*		*Ineff*		*AC*	*Ineff*
	C	*NC*	*C*	*NC*	*C*	*NC*	*C*	*NC*	*C*	*C*
$\boldsymbol{\beta}_1^G$	.24	.25	2.3	17.0	.15	.24	1.2	4.0	.05	1.1
	.05	.52	1.0	17.7	.04	.04	1.4	1.0	.05	1.1
	.05	.40	1.1	5.1	.04	−.03	1.0	1.0	−.07	1.0
$\boldsymbol{\beta}_2^G$	.06	.19	3.3	4.7	.09	.05	1.1	2.3	.10	1.2
	.07	.25	2.0	4.5	.04	.01	1.0	1.0	.00	1.0
	−.01	.81	1.7	34.2	−.02	.01	1.0	1.0	.02	1.0
$\delta_{1,1}$	.27	.30	1.0	2.1	.33	.41	1.0	3.4	.51	2.3
$\delta_{1,2}$	.21	.17	1.0	1.1	.20	.28	1.0	1.0	.47	3.0
$\delta_{1,3}$	.51	.53	2.3	1.0	.46	.55	2.0	1.0	.71	2.9
$\delta_{2,1}$	.47	.54	1.8	2.7	.52	.50	1.0	4.3	.50	2.3
$\delta_{2,2}$	.51	.50	1.0	1.4	.49	.46	1.1	2.8	.49	1.5
$\delta_{2,3}$	.25	.26	1.0	1.0	.36	.22	1.0	1.0	.22	1.5
$\boldsymbol{\eta}$	.11	.09	1.0	5.9	.10	.18	1.0	2.2	.03	1.1
σ_ε^2	.37	.72	5.0	1.6	.70	.70	4.4	5.9	.07	1.2

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

2.2 Sampling From Partially Marginalized Conditionals

We demonstrated in the last section that choosing the wrong parameterization may cause slow convergence for a sampler that is based on sampling fixed effects $\boldsymbol{\alpha}$ and the heterogeneous effects $\boldsymbol{\beta}^N$ from the full conditionals. Rather than trying to find the right parameterization for the case study at hand, one could apply a sampler that is insensitive to the parameterization. Note that a suitable parameterization does not exist for the case of $K > 1$, where at least one group is close to homogeneity and at least one other group is characterized by random-effects variances that dominate the observation error variance.

Common methods of constructing such a sampler are blocking, such that parameters are sampled jointly, and collapsing, where full conditional densities are substituted by marginal densities. The latter are obtained from integrating out part of the conditioning parameters.

The experiences reported for a normal linear mixed model by Chib and Carlin (1999) and Frühwirth-Schnatter and Otter (1999) suggest that it is desirable to group the fixed effects and the group-specific means and to sample them from the density where the random effects are integrated out. Furthermore, from the result reported by Gerlach, Carter, and Kohn (2000), we expect that it is also desirable to sample the group indicators from the density where the random effects are integrated out. This leads to the following sampler:

Algorithm 2 (Sampling from partially marginalized conditionals).

(i-2) Sample $\mathbf{S}^N$ from $\pi(\mathbf{S}^N|\boldsymbol{\phi}, \mathbf{y}^N)$.

(ii-2) Sample $\boldsymbol{\eta}$ from $\pi(\boldsymbol{\eta}|\mathbf{S}^N)$.

(iii-2) Sample $\boldsymbol{\phi}_L$, the fixed effects and the group-specific means, and $\boldsymbol{\beta}^N$, the random effects, conditional on $\mathbf{S}^N, \boldsymbol{\phi}_V$, and $\mathbf{y}^N$ from the joint normal distribution $\pi(\boldsymbol{\phi}_L, \boldsymbol{\beta}^N|\mathbf{S}^N, \boldsymbol{\phi}_V, \mathbf{y}^N)$. This step is carried out in two subblocks:

(iii-2a) Sample $\boldsymbol{\phi}_L$ from $\pi(\boldsymbol{\phi}_L|\mathbf{S}^N, \boldsymbol{\phi}_V, \mathbf{y}^N)$.

(iii-2b) Sample $\boldsymbol{\beta}^N$ from $\pi(\boldsymbol{\beta}^N|\mathbf{S}^N, \boldsymbol{\phi}_L, \boldsymbol{\phi}_V, \mathbf{y}^N)$.

(iv-2) Sample $\boldsymbol{\phi}_V$ from $\pi(\boldsymbol{\phi}_V|\boldsymbol{\beta}^N, \mathbf{S}^N, \boldsymbol{\phi}_L, \mathbf{y}^N)$.

Note that only steps (i) and (iii) differ between Algorithms 1 and 2. In step (i-2), we use a collapsed sampler for $\mathbf{S}^N$, where instead of sampling from the full conditional density $\pi(\mathbf{S}^N|\boldsymbol{\beta}^N, \boldsymbol{\phi}, \mathbf{y}^N)$, we sample from the marginal conditional density $\pi(\mathbf{S}^N|\boldsymbol{\phi}, \mathbf{y}^N)$, where the random effects $\boldsymbol{\beta}^N$ are integrated out. In step (iii-2) we use a blocked sampler for jointly sampling $\boldsymbol{\alpha}, \boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_K^G, \boldsymbol{\beta}^N$ rather than sampling $\boldsymbol{\alpha}$, $(\boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_K^G)$, and $\boldsymbol{\beta}^N$ in three different blocks conditional on the other parameters. Step (iii-2a) amounts to sampling the location parameters $\boldsymbol{\phi}_L = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_K^G)$ from the conditional density $\pi(\boldsymbol{\phi}_L|\mathbf{S}^N, \boldsymbol{\phi}_V, \mathbf{y}^N)$, where again the random effects $\boldsymbol{\beta}^N$ are integrated out.

To sample from the marginal conditional posterior densities in Algorithm 2, we derive the marginal model from (1), where the random effects are integrated out. Conditional on $\mathbf{S}^N$, the mixture of random-effects model (1) is a single mixed-effects model with heterogeneity in the variance of the random effects:

$$\mathbf{y}_i = \mathbf{X}_i^1\boldsymbol{\alpha} + \mathbf{X}_i^2\boldsymbol{\beta}_k^G + \mathbf{X}_i^2\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \qquad (12)$$

where, conditional on $S_i = k$, we obtain $\mathbf{b}_i = \boldsymbol{\beta}_i - \boldsymbol{\beta}_k^G \sim N(\mathbf{0}, \mathbf{Q}_k^G)$. If we integrate out $\mathbf{b}_i$, then we obtain

$$\mathbf{y}_i = \mathbf{X}_i^1\boldsymbol{\alpha} + \mathbf{X}_i^2\boldsymbol{\beta}_k^G + \boldsymbol{\varepsilon}_i^*, \qquad \boldsymbol{\varepsilon}_i^* \sim N\big(\mathbf{0}, \mathbf{X}_i^2\mathbf{Q}_k^G(\mathbf{X}_i^2)' + \sigma_\varepsilon^2\mathbf{I}\big). \qquad (13)$$

More details on steps (i-2), (iii-2), and (iv-2) are given in Appendix B. The following example demonstrates that this marginalization leads to a sampler that is insensitive to the selected parameterization of the mean structure.

Simulations 1 and 2 (Continued). We are now going to apply Algorithm 2 to Simulation 1 and 2 from the previous section.

From Algorithm 2's lag 1 sample autocorrelations and the inefficiency factors in Tables 1 and 2, we cannot make out any difference between the two methods of parameterization when sampling $\boldsymbol{\beta}_k^G$. Thus Algorithm 2 actually proves insensitive to the parameterization chosen for the mean structure. When we compare the values for $\boldsymbol{\beta}_k^G$ sampled by Algorithm 2 with the ones derived by Algorithm 1, we see that the partial marginalization leads to lower autocorrelations, as well as to improved efficiency.

2.3 Sampling From the Marginal Model

One special feature of Algorithm 2 is that the parameters $\boldsymbol{\phi}_L$ appearing in the mean structure, as well as the group indicators $\mathbf{S}^N$, are sampled from the marginal model, where the random effects are integrated out, whereas the variance parameters $\boldsymbol{\phi}_V$ are sampled from the full conditional model. The reason for this is that despite marginalization over the random effects, the conditional densities of the mean parameters $\boldsymbol{\phi}_L$ are from a well-known distribution family (multivariate normal), whereas the marginal distribution of the variance parameters $\pi(\boldsymbol{\phi}_V|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$ where the random effects are integrated out does not belong to a well-known family. However, the functional value of the nonnormalized marginal density $\pi^\star(\boldsymbol{\phi}_V|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$ is available from

$$\pi(\boldsymbol{\phi}_V|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N) \propto L(\mathbf{y}^N|\boldsymbol{\phi}_V, \boldsymbol{\phi}_L, \mathbf{S}^N)\pi(\boldsymbol{\phi}_V), \qquad (14)$$

where

$$L(\mathbf{y}^N|\boldsymbol{\phi}_V, \boldsymbol{\phi}_L, \mathbf{S}^N) = \prod_{i=1}^N L\big(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}_{S_i}^G, \mathbf{Q}_{S_i}^G, \sigma_\varepsilon^2\big). \qquad (15)$$

From the marginal model (13), we obtain that the likelihood $L(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}_k^G, \mathbf{Q}_k^G, \sigma_\varepsilon^2)$ is the likelihood of the normal distribution $N(\mathbf{y}_i; \mathbf{X}_i^1\boldsymbol{\alpha} + \mathbf{X}_i^2\boldsymbol{\beta}_k^G, \mathbf{X}_i^2\mathbf{Q}_k^G(\mathbf{X}_i^2)' + \sigma_\varepsilon^2\mathbf{I})$.

Because the nonnormalized density $\pi^\star(\boldsymbol{\phi}_V|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$ is available in closed form, we use the Metropolis–Hastings algorithm to sample $\boldsymbol{\phi}_V$ from $\pi(\boldsymbol{\phi}_V|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$. Details are given in Appendix C, where we use a mixture of inverted Wishart densities as a proposal density.

Algorithm 3 (Sampling from fully marginalized conditionals).

(i-3) Sample $\mathbf{S}^N$ from $\pi(\mathbf{S}^N|\boldsymbol{\phi}, \mathbf{y}^N)$.

(ii-3) Sample $\boldsymbol{\eta}$ from $\pi(\boldsymbol{\eta}|\mathbf{S}^N)$.

(iii-3) Sample all fixed effects, random effects, and group-specific means $\boldsymbol{\phi}_L$, $\boldsymbol{\beta}^N$ conditional on $\mathbf{S}^N$, $\boldsymbol{\phi}_V$, $\mathbf{y}^N$ from the joint normal distribution $\pi(\boldsymbol{\phi}_L, \boldsymbol{\beta}^N|\mathbf{S}^N, \boldsymbol{\phi}_V, \mathbf{y}^N)$, as in (iii-2).

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

(iv-3) Sample the variance parameters $\boldsymbol{\phi}_V$ conditional on $\boldsymbol{\phi}_L$, $\mathbf{S}^N$, $\mathbf{y}^N$, where the random effects are integrated out:

(iv-3a) For $k = 1, \ldots, K$, sample $\mathbf{Q}_k^G$ from $\pi(\mathbf{Q}_k^G | \mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$, where $\mathbf{Q}_{-k}^G$ denotes all covariance matrices but $\mathbf{Q}_k^G$.

(iv-3b) Sample σ_ε^2 from $\pi(\sigma_\varepsilon^2 | \mathbf{Q}_1^G, \ldots, \mathbf{Q}_K^G, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$.

Simulations 1 and 2 (Continued). We applied Algorithm 3 to our simulated data. Because of the invariance to the parameterization (see Sec. 2.2), we use only the centered parameterization here. To compare the performance of this algorithm with the performance of Algorithms 1 and 2, we listed the autocorrelations at lag 1 and the inefficiency factors in Tables 1 and 2. We carry out 3,000 iterations and keep the last 2,000 for estimation. We see that the Metropolis–Hastings step did not lead to better behavior of the sampler for the group-specific variances for either Simulation 1 or Simulation 2. Only the model error variance σ_ε^2 may be sampled more efficiently by the Metropolis–Hastings step.

3. DEALING WITH UNIDENTIFIABILITY

3.1 About Unidentifiability and Label Switching

In what follows we use the notation $\boldsymbol{\psi} = (\boldsymbol{\beta}^N, \mathbf{S}^N, \boldsymbol{\phi})$ to denote all unknown parameters and $\mathcal{A}$ to denote the unconstrained parameter space. The unconstrained parameter space $\mathcal{A}$ consists of $K!$ disjunct subspaces $\mathcal{L}_1, \ldots, \mathcal{L}_{K!}$, differing only in the way of labeling the groups, $\mathcal{A} = \bigcup_{l=1}^{K!} \mathcal{L}_l$. To each labeling subspace $\mathcal{L}_l$ corresponds a certain permutation ρ_l of $\{1, \ldots, K\}$ telling us with which of the K groups the various group-specific components of a parameter $\boldsymbol{\psi} \in \mathcal{A}$ have to be associated. Without loss of generality, we may assume that $\rho_1(\cdot)$ is equal to the identity. Therefore, if $\boldsymbol{\psi} \in \mathcal{L}_1$, then for each $k = 1, \ldots, K$, the components $(\boldsymbol{\beta}_k^G, \mathbf{Q}_k^G, \eta_k)$ of $\boldsymbol{\phi}$ correspond to group k. In $\mathcal{L}_l$ with $l > 1$, however, $(\boldsymbol{\beta}_k^G, \mathbf{Q}_k^G, \eta_k)$ correspond to group $\rho_l(k)$ rather than k as before. This label switching between the labeling subspaces causes the unconstrained posterior to have multiple, at most $K!$, modes. The modes are equivalent if the prior $\pi(\boldsymbol{\phi})$ is invariant to relabeling the groups.

For Bayesian estimation via MCMC methods, it is essential to produce draws from a unique labeling subspace, if interest lies in estimation of group-specific parameters such as $\boldsymbol{\beta}_k^G$, $\mathbf{Q}_k^G$, η_k or classification probabilities $\Pr(S_i = k|\mathbf{y}^N)$. Otherwise, label switching could be present, rendering meaningless any estimation of group-specific parameters from the MCMC output. There is a geometric aspect to the identification of a unique labeling in that simulations are confined to only *one* of the $K!$ possible modal regions.

It is important to emphasize that identifying a unique labeling is different from formal identifiability achieved by introducing a constraint, $\mathcal{R}$. In general, an identifiability constraint is defined in terms of a subset $\mathcal{R}$ of the unrestricted parameter space $\mathcal{A}$, such that for all parameters $\boldsymbol{\psi} \in \mathcal{A}$, $\rho(\boldsymbol{\psi}) \in \mathcal{R}$ for exactly one permutation $\rho_{\boldsymbol{\psi}}(\cdot)$ (see Stephens 1997, p. 43). An example would be the following constraint on the weights:

$$\mathcal{R}_{\boldsymbol{\eta}} : \eta_1 < \cdots < \eta_K, \qquad (16)$$

which was applied by Aitkin and Rubin (1985) and Lenk and DeSarbo (2000) as a standard constraint for this type of mixture models. But this constraint may be an order relation for a single component of the group-specific parameter or may involve more than one component. With an identifiability constraint $\mathcal{R}$, formal identifiability is achieved in the sense that if two parameters $\boldsymbol{\phi}$ and $\tilde{\boldsymbol{\phi}}$ define the same probability law for all possible observations $\mathbf{y}^N$, then $\boldsymbol{\phi}$ and $\tilde{\boldsymbol{\phi}}$ need to be the same:

$$L(\mathbf{y}^N|\boldsymbol{\phi}) = L(\mathbf{y}^N|\tilde{\boldsymbol{\phi}}) \quad \longrightarrow \quad \boldsymbol{\phi} = \tilde{\boldsymbol{\phi}}. \qquad (17)$$

$L(\mathbf{y}^N|\boldsymbol{\phi})$ is the marginal likelihood, where $(\boldsymbol{\beta}^N, \mathbf{S}^N)$ are integrated out.

It is a common misunderstanding that formal identifiability through an *arbitrary* identifiability constraint $\mathcal{R}$ leads to a unique labeling. This is not necessarily the case. An identifiability constraint restricts the parameter space $\mathcal{A}$ to a subspace in which no two different parameters define the same probability law to achieve (17). But the parameters in this subspace need not come from one labeling subspace—or, in more geometric terms, from one modal region. As a consequence, if Bayesian analysis is carried out on that subspace defined by the identifiability constraint $\mathcal{R}$, then label switching might still be present. An identifiability constraint will lead to a unique labeling only if it respects the geometry of the posterior and is able to separate the modes of the unconstrained posterior.

For illustration, we simulated 200 datasets, each containing $N = 100$ bivariate observations from the four-component mixture model

$$\mathbf{y}_t \sim \eta_1 N(\mathbf{y}_t; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \eta_2 N(\mathbf{y}_t; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) + \eta_3 N(\mathbf{y}_t; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}) + \eta_4 N(\mathbf{y}_t; \boldsymbol{\mu}_4, \boldsymbol{\Sigma}), \qquad (18)$$

where $\boldsymbol{\eta} = (.1\ .24\ .26\ .4)'$, $\boldsymbol{\mu}_1 = (-2\ -2)'$, $\boldsymbol{\mu}_2 = (-2\ 2)'$, $\boldsymbol{\mu}_3 = (2\ -2)'$, $\boldsymbol{\mu}_4 = (2\ 2)'$, and $\boldsymbol{\Sigma} = \sigma^2 \cdot \mathbf{I}$ with $\sigma^2 = .01$ and the identity matrix $\mathbf{I}$. This is a simple special case of model (1) with $\mathbf{Q}_1^G = \cdots = \mathbf{Q}_K^G = \mathbf{0}$, $\mathbf{X}_i^1 = \mathbf{0}$, $\mathbf{X}_i^2 = \mathbf{I}$, and $\boldsymbol{\beta}_k^G = \boldsymbol{\mu}_k$.

To illustrate the difference between the unconstrained and the constrained posterior, Figure 1 shows simulations from various bivariate marginal densities, $\pi(\mu_{k,1}, \mu_{k,2}|\mathbf{y}^N)$, $k = 1, \ldots, 4$, for one randomly selected dataset. All simulations are based on the prior distributions $\boldsymbol{\mu}_j \sim N(\mathbf{0}, 4 \cdot \mathbf{I})$, $\sigma^2 \sim IG(2, .01)$ and $\boldsymbol{\eta} \sim D(4, \ldots, 4)$ and were obtained by permutation sampling (see Frühwirth-Schnatter 2001a). The first row shows simulations from the unconstrained posterior, which obviously has multiple modes and is invariant to permutations of the labels (all four figures are identical). The second row shows simulations from the constrained posterior under the constraint (16) on the weights, $\mathcal{R}_{\boldsymbol{\eta}} : \eta_1 < \cdots < \eta_4$. Obviously the simulations exhibit label switching. Because two of the groups have weights that, although different, are close to each other, we achieve formal identification by constraint (16) but no unique labeling. From the simulations shown in the first row of Figure 1, a constraint that separates the four modal regions can be easily found. One possibility is the following constraint on the means:

$$\mathcal{R}_2 : \max(\mu_{1,1}, \mu_{2,1}) < \min(\mu_{3,1}, \mu_{4,1}), \qquad \mu_{1,2} < \mu_{2,2}, \mu_{3,2} < \mu_{4,2}. \qquad (19)$$

To emphasize the difference between formal identification within a likelihood analysis and unique labeling within a

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

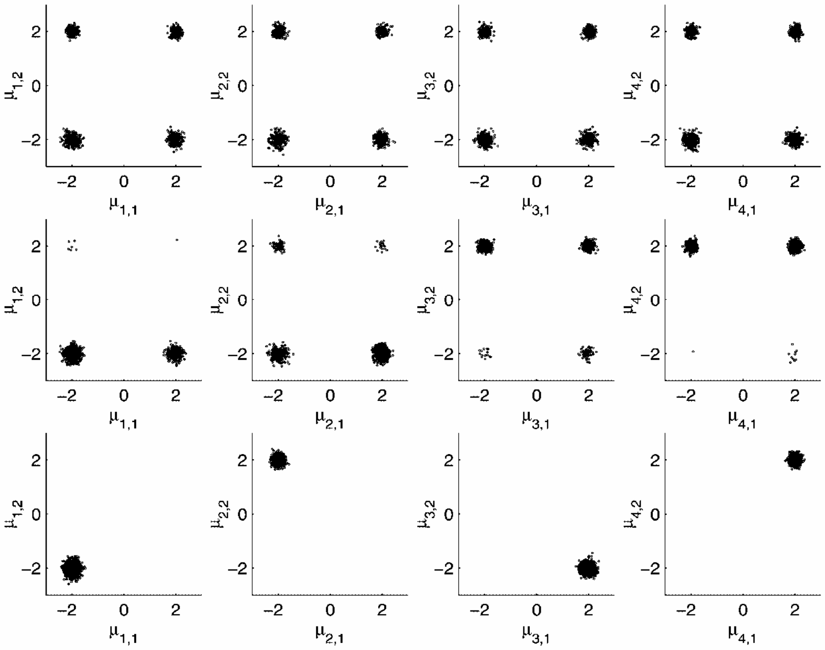

Figure 1. Simulations From the Various Bivariate Marginal Densities $\pi(\mu_{k,1}, \mu_{k,2}|\mathbf{y}^N)$ for $K = 4$. First row, random permutation sampling from the unconstrained posterior; second row, permutation sampling under constraint (16); last row, permutation sampling under constraint (19).

Bayesian analysis, the last row in Figure 1 shows simulations from the various bivariate marginal densities $\pi(\mu_{k,1}, \mu_{k,2}|\mathbf{y}^N)$ under constraint (19). In contrast with constraint (16), this constraint induces a unique labeling.

The fact that formal identification by constraint (16) does not necessarily induce a unique labeling remains undetected if a MCMC sampler is used for constrained estimation that sticks at the current labeling subspace. An example is the slice sampler discussed by Lenk and DeSarbo (2000). As can be seen from comparing the first and second rows of Figure 2, the draws from slice sampling under constraint (16) stay within one labeling subspace only, missing part of the parameter space constrained to $\mathcal{R}_{\boldsymbol{\eta}}$. This might introduce a bias toward the constraint in comparison to draws obtained under constraint (19), which induces a unique labeling.

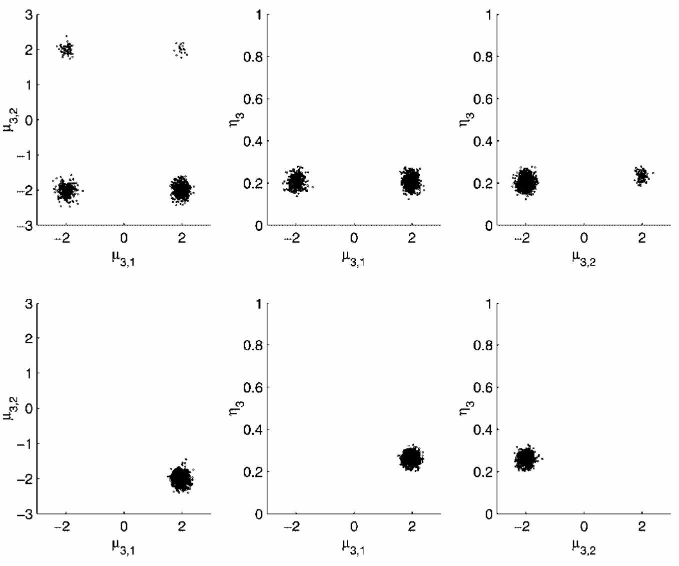

Figure 2. Simulations From the Various Bivariate Marginal Densities $\pi(\mu_{3,1}, \mu_{3,2}|\mathbf{y}^N)$, $\pi(\mu_{3,1}, \eta_3|\mathbf{y}^N)$ and $\pi(\mu_{3,2}, \eta_3|\mathbf{y}^N)$ for $K = 4$. First row, permutation sampling under constraint (16); second row, slice sampling under constraint (16).

The influence of the constraint is evident from Figure 3, in which we compare the posterior estimates of $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ obtained from permutation sampling under constraint (19) with slice sampling under constraint (16) for all 200 simulated datasets. Whereas a bias toward the constraint $\eta_2 < \eta_3$ is introduced for η_2 and η_3, the constraint causes less variation of η_1 and η_4 around the true values, and has no effect on estimation of $\boldsymbol{\mu}$.

Whether or not this bias matters depends on what inference we draw from the simulations. (See Frühwirth-Schnatter 1999 for discussion on the considerable effect of this bias on estimates of the model likelihood.) Assume in the present context that we want to estimate the difference $\mathbf{D}$ between various group sizes,

$$\mathbf{D} = \begin{pmatrix} \eta_2 - \eta_1 \\ \eta_3 - \eta_2 \\ \eta_4 - \eta_3 \end{pmatrix}, \tag{20}$$

with the true value being $\mathbf{D} = (.14, .02, .14)$. If we want to obtain an estimator with small systematic error (i.e., small bias $\hat{\mathbf{D}} - \mathbf{D}$), then the bias introduced by constraint $\mathcal{R}_{\boldsymbol{\eta}}$ of course matters (Table 3). If we are interested in obtaining an (possi-

Table 3. Comparing Slice Sampling Under Constraint (16) and Permutation Sampling Under Constraint (16)

		$\eta_2 - \eta_1$	$\eta_3 - \eta_2$	$\eta_4 - \eta_3$
Bias	Slice sampling	−.0394	.0368	−.0229
	Permutation sampling	−.0182	−.00555	−.0159
Mean squared error	Slice sampling	.00261	.00190	.00285
	Permutation sampling	.00314	.00394	.00479

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

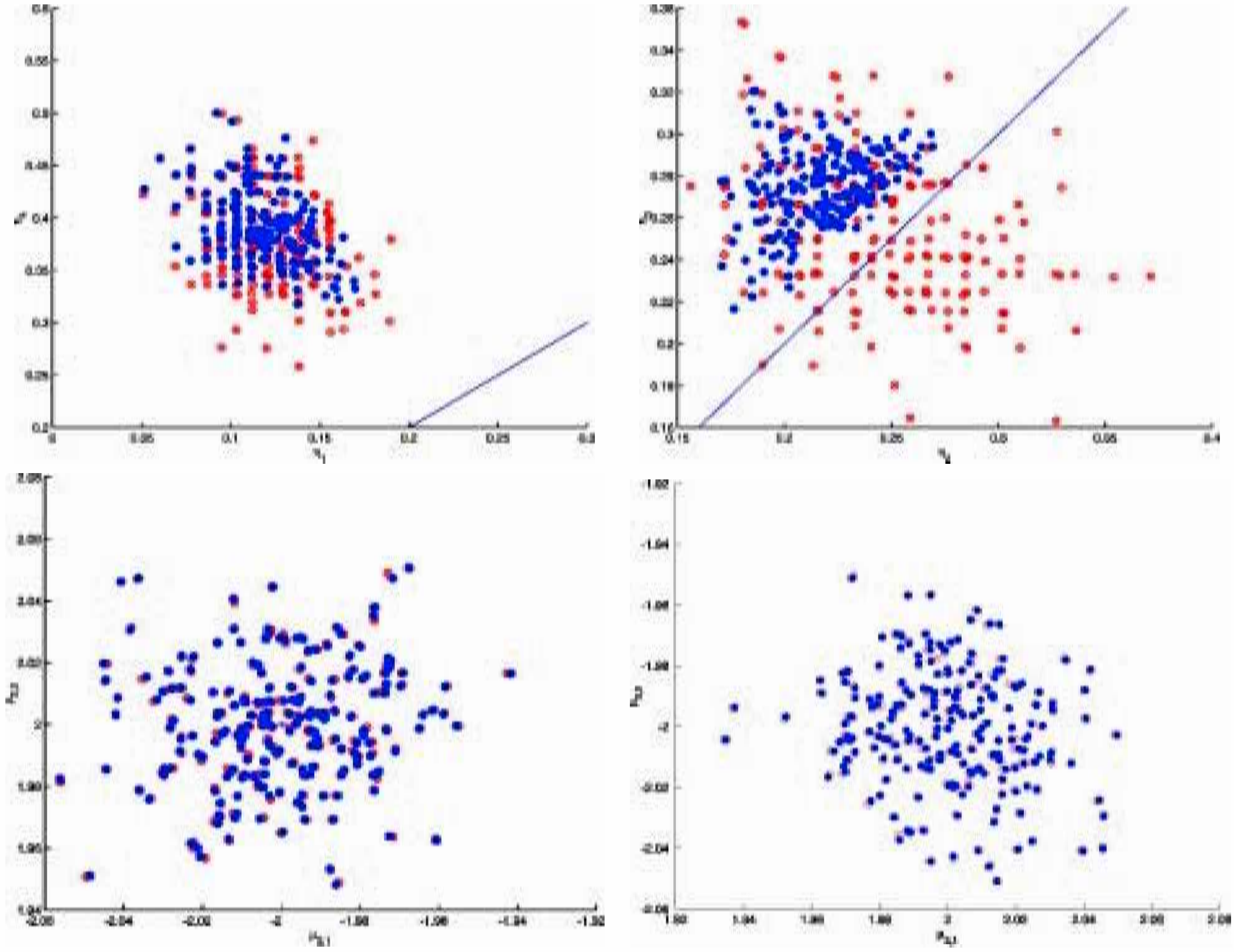

Figure 3. Posterior Estimates Obtained for 200 Datasets Simulated From Model (18). ⊗, permutation sampling under constraint (19); •, slice sampling under constraint (16); $+$, true value.

bly biased) estimator with small variance around the true value [i.e., small mean squared error $(\hat{\mathbf{D}}-\mathbf{D})^2$], then the bias does not matter as long as it is small compared with the variation of the estimator. For the present case study, the biased estimator has even smaller mean squared error than the estimator obtained from constraint (19) (see Table 3).

3.2 Statistical Inference From the Unidentified Model

We emphasize that for many estimation problems arising in the empirical analysis of the mixtures of random effects models, identifying a unique labeling is not necessary. This is the case if we want to estimate a functional of the augmented parameter vector $f(\boldsymbol{\psi})$ that is invariant to relabeling the indices. For such a functional, the expectation with respect to posterior simulations constrained to any of the unique labeling subspaces is equal to the expectation with respect to posterior simulations from the unconstrained space $\mathcal{A}$,

$$E_{\mathcal{L}_l}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big)=E_{\mathcal{L}_s}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big)=E_{\mathcal{A}}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big). \qquad (21)$$

A proof of (21) is given in Appendix D. A practical consequence of (21) is that many quantities may be estimated without introducing a unique labeling. Among these quantities are obviously parameters that are common to all components of the mixture, such as $\boldsymbol{\alpha}$ and σ_ε^2, and, notably, the *individual parameters* $\boldsymbol{\beta}_i$, because

$$\boldsymbol{\beta}_i \sim \sum_{k=1}^{K}\eta_k N(\boldsymbol{\beta}_k^G,\mathbf{Q}_k^G)=\sum_{k=1}^{K}\eta_{\rho(k)}N\big(\boldsymbol{\beta}_{\rho(k)}^G,\mathbf{Q}_{\rho(k)}^G\big).$$

Consequently, the behavior of each subject under designs $\mathbf{X}_i^{1,\star}$, $\mathbf{X}_i^{2,\star}$ different from that used for estimation may be predicted without a unique labeling. Further example for quantities that are invariant to relabeling are all moments of the distribution of heterogeneity, for example, the mean or the covariance matrix

$$\boldsymbol{\alpha}_{\boldsymbol{\beta}}=\sum_{k=1}^{K}\boldsymbol{\beta}_k^G\eta_k, \qquad \mathbf{Q}=\sum_{k=1}^{K}\big(\boldsymbol{\beta}_k^G(\boldsymbol{\beta}_k^G)'+\mathbf{Q}_k^G\big)\eta_k-\boldsymbol{\alpha}_{\boldsymbol{\beta}}\boldsymbol{\alpha}_{\boldsymbol{\beta}}'.$$

4. AN EMPIRICAL STUDY FROM METRIC CONJOINT ANALYSIS

4.1 The Data and the Model

The data come from a brand–price trade-off study in the mineral water category. Each of 213 Austrian consumers stated their likelihood of purchasing 15 different product profiles offering five brands of mineral water—Römerquelle (RQ),

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

Vöslauer, Juvina, Waldquelle, and one brand not available in Austria, Kronsteiner—at three different prices [2.80, 4.80, and 6.80 Austrian shillings (ATS)] on 20-point rating scales (where higher values indicate a greater likelihood of purchasing). In an attempt to make the full brand-by-price factorial less obvious to consumers, the price levels varied in the range of ±.1 ATS around the respective design levels such that mean prices of brands in the design were not affected (Elrod, Louviere, and Davey 1992).

We used a fully parameterized matrix, $\mathbf{X}_i^2$, with 15 columns corresponding to the constant, four brand contrasts, a linear and a quadratic price effect, and four brand–linear price and four brand–quadratic price interaction effects. We used dummy coding for the brands. The unknown brand, Kronsteiner, was chosen as the baseline. We subtracted the smallest price from the linear price column in matrix $\mathbf{X}_i^2$ and computed the quadratic price contrast from the centered linear contrast.

4.2 Identifying a Unique Labeling: Exploratory Evaluation and Selection of Constraints

We demonstrated in Section 3.1 that an arbitrary identifiability constraint does not guarantee a unique labeling. The ideas described by Frühwirth-Schnatter (2001a) may be applied to explore simulations from the unconstrained posterior to find an identifiability constraint $\mathcal{R}$ that can separate the posterior modes, and thus to identify simulations from one labeling subspace. In Section 3.1 we distinguished between formal identifiability and identification of a unique labeling subspace. To find such a unique subspace, we have to respect the geometry of the posterior. We recommend looking at different graphical presentations (e.g., marginal density plots and two- or three-dimensional scatterplots) of posterior simulations of group-specific parameters to learn about the geometry of the posterior. Appropriate constraints are, of course, less obvious the higher the number of groups K in the model and the greater the dimensionality of the group-specific parameter vectors involved. Our experience from many applications, some of them involving a large number of groups K and high-dimensional parameter vectors, suggests that as long as the data supports the number of groups chosen for the model, a graphical inspection of the posterior will lead to constraints that identify a unique labeling. When there are many groups, the identification of a unique labeling proceeds sequentially. Usually, there is one parameter or an easily identified linear combination of very few parameters that divides the posterior into two subsets containing several groups each. Conditional on this first restriction, each subset is investigated in turn for an additional restriction that identifies even smaller subsets of groups until finally every group is identified. Note, however, that any suitable constraint is only an indirect devise to identify a unique labeling, and therefore is not necessarily unique. (For more details, see the case studies in Frühwirth-Schnatter 2001a,b, and Kaufmann and Frühwirth-Schnatter 2002.) For an identification problem that was more challenging due to the higher number of classes, Otter, Tüchler, and Frühwirth-Schnatter (2002) identified a latent class model with nine classes for the present case study. Alternative approaches to identify a unique labeling have been discussed by Celeux et al. (2000) and Stephens (2000).

Tüchler, Frühwirth-Schnatter, and Otter (2002) reported that Bayesian model choice criteria indicated that the optimal model in the model class with homogeneous error variances and fixed quadratic price interactions has three groups. For this model, we explore simulations from the unconstrained posterior to find an (not necessarily unique) identifiability constraint (Fig. 4). From this plot, we can readily identify the three simulation clusters and can easily make out an identifiability constraint, namely $\mathcal{R}$: $\text{price}_1 < \min(\text{price}_{2,3})$, to separate the first group from the other two, and $\text{RQ}_2 > \text{RQ}_3$ to distinguish between the two remaining groups. The simulations of the constrained posterior resulting when applying this constraint are plotted in Figure 5.

Figure 6 shows the posterior simulations for a model with $K = 4$ groups and fixed quadratic price interactions. From this plot, it is obvious that the data support only three simulation clusters. The fourth group is spread over the parameter space, and therefore this plot can be used as an empirical indicator for determining the optimal number of groups $K = 3$.

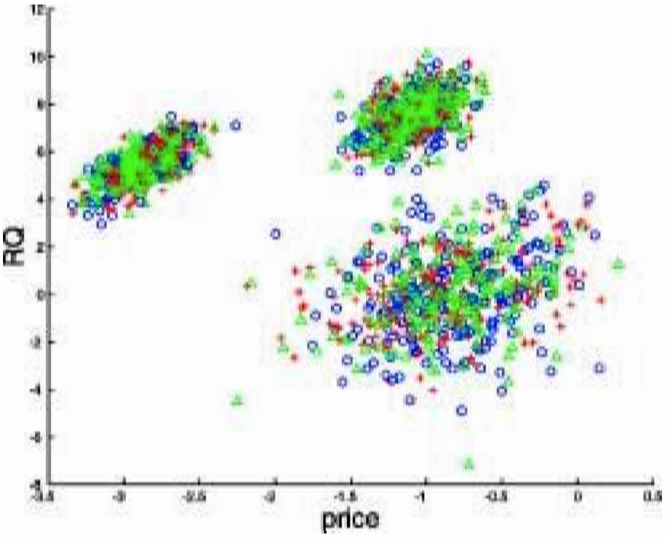

Figure 4. Modeling the Empirical Data by a Three-Groups Model; Simulations of the Group-Specific Means From the Unconstrained Posterior.

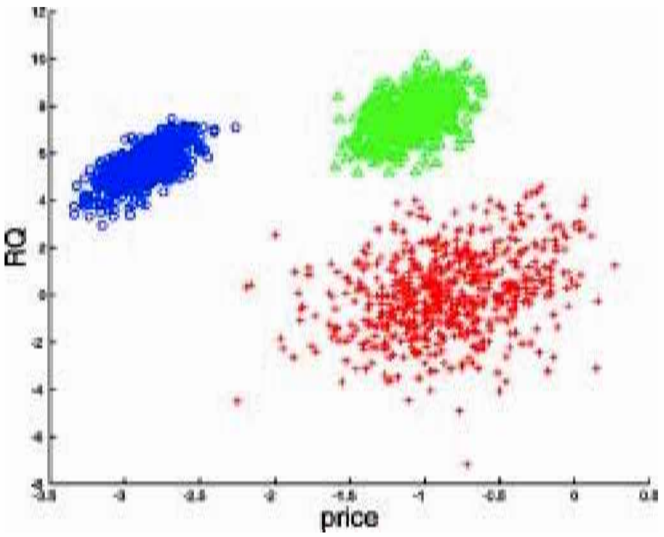

Figure 5. Modeling the Empirical Data by a Three-Group Model; Simulations of the Group-Specific Means From the Constrained Posterior.

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

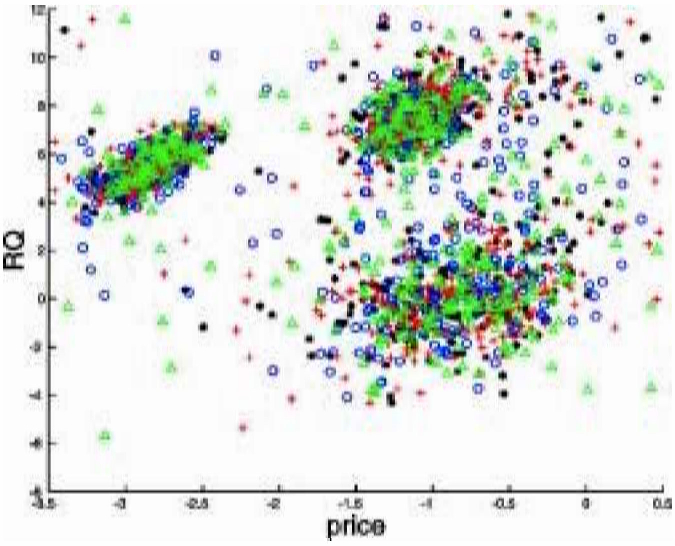

Figure 6. Modeling the Empirical Data by a Four-Group Model; Simulations of the Group-Specific Means From the Unconstrained Posterior.

4.3 Comparison of the Three Algorithms and the Two Parameterizations

To investigate the influence of the two types of parameterization on Algorithm 1 (full conditional Gibbs sampling) and on Algorithm 2 (partially marginalized Gibbs sampling), we choose a model with two groups and fixed quadratic price interaction effects. We did not use the three-group model with fixed quadratic interactions in this comparison because Algorithm 1's behavior in combination with the noncentered parameterization is extraordinarily bad in this case. It was not possible to sample the optimal three-group model with fixed quadratic interactions with this algorithm. This is illustrated in Figure 7, which shows marginal densities of the group weights for the optimal three-group model. Figure 7a indicates that Algorithm 1 (centered), Algorithm 2, and Algorithm 3 estimate group sizes η of approximately .47, .43, and .1. In sharp contrast, we see from Figure 7b that Algorithm 1 with the *noncentered* parameterization samples two groups of size .59 and .41 and a third group of size close to 0. This has a considerable influence on model estimation. If we estimated our empirical data just with Algorithm 1 in combination with the noncentered parameterization the marginal densities of the group weights plotted on the right side of Figure 7 would suggest that a three-group model is a model with too high a number of groups, and we would end up with a model with fewer groups.

For the two-group model with fixed quadratic interactions the MCMC samples obtained from Algorithms 1 and 2 under the two different parameterizations are again compared by their lag 1 sample autocorrelations and the inefficiency factors. The values are listed in Table 4 for the group-specific means $\boldsymbol{\beta}_k^G$ and fixed effects $\boldsymbol{\alpha}$. We see that Algorithm 2 behaves invariantly with regard to the parameterization. For Algorithm 1, the behavior is two fold. For the group-specific means $\boldsymbol{\beta}_k^G$, the centered parameterization is preferred, whereas for the fixed effects $\boldsymbol{\alpha}$, the noncentered parameterization turns out to be better, so we cannot decide which parameterization to choose. In addition, the results of Table 4 clearly advocate for choosing Algorithm 2, because nearly all autocorrelations and inefficiency factors lead to better values for Algorithm 2 than for Algorithm 1 regardless of the parameterization.

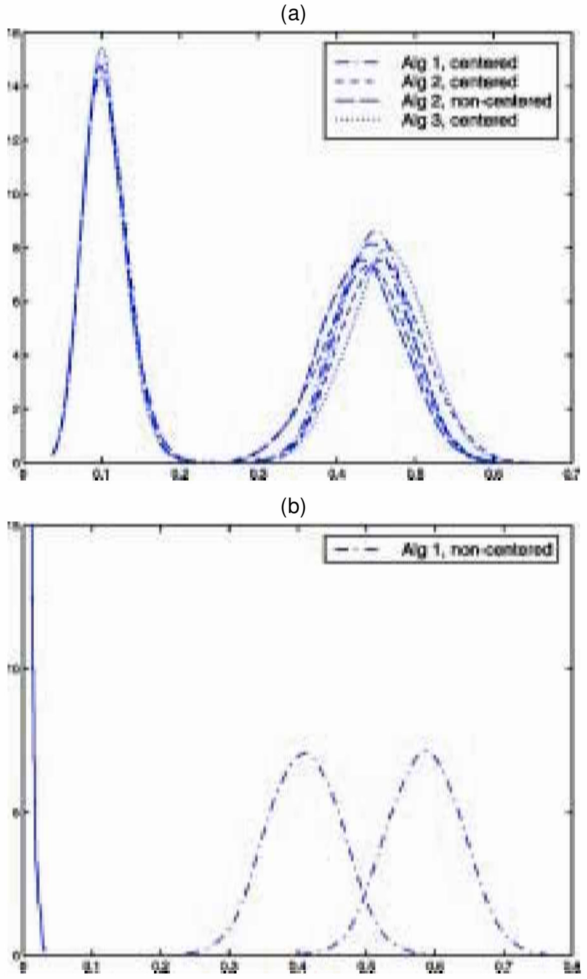

Figure 7. Marginal Densities of the Group-Weights η When Applying the Model With Three Groups and Fixed Quadratic Interactions to the Empirical Data. [(a)-· -· -· -, Algorithm 1, centered; - - - - -, Algorithm 2, centered; ———–, Algorithm 2, noncentered; · · · · · ·, Algorithm 3, centered. (b) -· -· -· -, Algorithm 1, noncentered.]

As in the simulation studies, it again turned out that applying Algorithm 3's Metropolis–Hastings steps to sample the group-specific covariances $\mathbf{Q}_1^G, \ldots, \mathbf{Q}_K^G$ does not improve the sampler's autocorrelation and efficiency in comparison to Algorithm 2. However, sampling the model error variance σ_ε^2 with Algorithm 3's Metropolis–Hastings steps improves the quality of the sampler. The autocorrelation at lag 1 is .62 for Algorithm 2 and .16 for Algorithm 3, and the inefficiency factor decreases from 6.5 to 1.0.

5. CONCLUDING REMARKS

In this article we tried to provide a deeper understanding of the circumstances under which full conditional Gibbs sampling of the heterogeneity model will be a sensible tool for Bayesian estimation. It turned out that full conditional Gibbs sampling is sensitive to the parameterization used for the mean, and we demonstrated that the best thing one could do for a model with a *normal* observation density is to use an at least partly marginalized sampler where the random effects are integrated out.

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

Table 4. Empirical Study, Model With Two Groups and Fixed Quadratic Interactions, Autocorrelation at Lag 1 (AC) and Inefficiency Factors (Ineff) for Algorithms 1 and 2 With the Centered (C) and the Noncentered (NC) Parameterization

	Algorithm 1				*Algorithm 2*			
	AC		*Ineff*		*AC*		*Ineff*	
	C	*NC*	*C*	*NC*	*C*	*NC*	*C*	*NC*
β_1^G	.26	.75	3.1	74.5	.15	.15	2.8	4.4
	.35	.85	14.4	97.5	.25	.25	4.4	10.4
	.30	.88	23.5	111.7	.29	.21	4.7	8.0
	.50	.68	70.9	12.1	.46	.36	14.7	8.4
	.43	.60	9.5	4.5	.21	.18	5.2	5.8
	.43	.52	4.2	3.0	.23	.24	1.0	3.9
	.57	.42	5.5	14.8	.06	.06	1.7	3.1
	.63	.58	10.7	7.9	.19	.19	3.1	1.9
	.59	.65	15.0	47.6	.16	.15	1.6	2.3
	.58	.61	1.9	3.7	.14	.18	2.5	2.1
	.65	.55	2.1	2.2	.09	.09	1.6	1.6
β_2^G	.24	.74	7.8	23.8	.18	.11	6.3	4.1
	.37	.65	8.1	29.7	.18	.14	3.1	2.3
	.34	.62	9.4	23.7	.17	.11	3.2	3.1
	.48	.70	29.2	24.3	.37	.30	7.7	9.5
	.28	.64	5.7	5.1	.10	.13	3.5	1.5
	.50	.81	70.9	58.1	.42	.31	10.0	8.4
	.42	.52	15.2	2.5	.07	.03	4.3	1.7
	.47	.68	14.4	32.8	.19	.19	5.7	6.6
	.53	.68	43.3	35.9	.20	.23	4.0	7.6
	.53	.66	7.6	35.7	.19	.20	5.4	7.7
	.50	.61	4.5	13.3	.08	.09	1.7	1.7
α	.82	.01	9.9	1.0	−.02	−.02	1.0	1.0
	.80	.02	15.6	1.0	.04	.00	1.4	1.0
	.80	.04	21.2	1.1	.01	.00	1.2	1.0
	.81	.02	10.8	2.0	.01	.01	1.0	1.0

Marginalization is usually not possible for models with nonnormal observation densities and understanding full conditional Gibbs sampling is even more important for this case. Although we did not investigate nonnormal models, we believe that the results presented here will generalize to this case.

A limitation of this article is that we studied only the effect of the parameterization of the mean on the performance of full conditional Gibbs sampling. An additional and equally important issue that we did not investigate is how parameterizing the variance structure influences the efficiency of the MCMC sampler. From Tables 1 and 2, we hypothesize that the current method for parameterizing the variance structure is rather efficient, if the variances are not too small. In the case of small variances using the parameterization that is centered in the mean around **0** ("noncentered" in the terminology of the article) helps improve the efficiency of estimating the mean parameters, but does not seem to improve the efficiency of variance parameters estimation. How to improve the estimation of variances in cases of small variances remains to be investigated, however. Meng and Van Dyck (1998) and Van Dyck and Meng (2001) explored a parameterization where the variance of the random effects is centered around the standard covariance matrix equal to the identity matrix **I** for random coefficient models. Extending their ideas to mixture of random-effects models yields

$$\mathbf{y}_i = \mathbf{X}_i^1\boldsymbol{\alpha} + \mathbf{X}_i^2\boldsymbol{\beta}_k^G + \mathbf{X}_i^2(\mathbf{Q}_k^G)^{1/2}\mathbf{z}_i + \boldsymbol{\varepsilon}_i, \qquad \text{if } S_i = k,$$

and

$$\mathbf{z}_i \sim N(\mathbf{0}, \mathbf{I}), \qquad i = 1, \ldots, N,$$

with the random effects centered around **I** instead of $\mathbf{Q}_{S_i}$. We leave the investigation of this parameterization to future research, however.

ACKNOWLEDGMENTS

This research was supported by the Austrian Science Foundation (FWF) under grant SFB 010 (Adaptive Information Systems and Modeling in Economics and Management Science).

APPENDIX A: THE PRIOR

The following prior information was used in the estimation:

- We impose the following prior on the effects for the centered parameterization. We assume that the group-specific effects $\boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_K^G$ are a priori independently identically distributed with $\boldsymbol{\beta}_k^G \sim N(\mathbf{c}_0, \mathbf{C}_0)$ for $k = 1, \ldots, K$, whereas the fixed effects $\boldsymbol{\alpha}$ are a priori normally distributed with $\boldsymbol{\alpha} \sim N(\mathbf{a}_0, \mathbf{A}_0)$. This prior must be transformed to obtain the prior under the noncentered parameterization.
- The group-specific covariances $\mathbf{Q}_1^G, \ldots, \mathbf{Q}_K^G$ are a priori independently identically distributed with $\mathbf{Q}_k^G \sim IW(\nu_0^Q, \mathbf{S}_0^Q)$.
- The probability distribution of $\boldsymbol{\eta}$ follows a Dirichlet prior, $D(e_{0,1}, \ldots, e_{0,K})$.
- The observation equation's variance σ_ε^2 is a priori inverted-gamma distributed with $IG(\nu_{\varepsilon,0}, S_{\varepsilon,0})$.

In the empirical study of Section 4 $\mathbf{c}_0$ and $\mathbf{a}_0$ are the ordinary least squares (OLS) estimates obtained by a model with

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

fixed effects only. $\mathbf{C}_0$ and $\mathbf{A}_0$ are diagonal matrices with diagonal elements equal to 25. We select $\nu_0^Q = 10$. $\mathbf{S}_0^Q$ is derived from the relationship $E(\mathbf{Q}_k^G) = (\nu_0^Q - (d+1)/2)^{-1}\mathbf{S}_0^Q$, where $E(\mathbf{Q}_k^G)$ was computed by individual OLS estimation and d is the dimensionality of $\mathbf{Q}_k^G$. We select $e_{0,k} = 1$ for $k = 1, \ldots, K$, which leads to a uniform prior on the unit simplex, and we stay fully noninformative about σ_ε^2.

APPENDIX B: DETAILS OF ALGORITHM 2

B.1 Sampling ($\boldsymbol{\phi}_L$,$\boldsymbol{\beta}^N$) by a Blocked Gibbs Sampler

We give details only for the centered parameterization. The conditional posterior of $\boldsymbol{\phi}_L$ and $\boldsymbol{\beta}^N$ partitions as

$$\begin{aligned}\pi(\boldsymbol{\phi}_L, \boldsymbol{\beta}^N|\mathbf{y}^N, \boldsymbol{\phi}_V, \mathbf{S}^N)&\\ &= \pi(\boldsymbol{\beta}^N|\boldsymbol{\phi}_L, \mathbf{y}^N, \boldsymbol{\phi}_V, \mathbf{S}^N)\pi(\boldsymbol{\phi}_L|\mathbf{y}^N, \boldsymbol{\phi}_V, \mathbf{S}^N)\\ &\propto \left(\prod_{i=1}^N \pi(\boldsymbol{\beta}_i|\boldsymbol{\phi}_L, \mathbf{y}_i, \boldsymbol{\phi}_V, S_i)\right)\pi(\boldsymbol{\phi}_L|\mathbf{y}^N, \boldsymbol{\phi}_V, \mathbf{S}^N).\end{aligned}$$

The conditional posterior of the random effects $\boldsymbol{\beta}_i$ is the normal distribution $\pi(\boldsymbol{\beta}_i|\mathbf{y}_i, \boldsymbol{\phi}_L, \boldsymbol{\phi}_V, S_i) \sim N(\boldsymbol{\beta}_{S_i}^G + \hat{\mathbf{b}}_i, \mathbf{B}_i)$ with

$$\hat{\mathbf{b}}_i = \mathbf{B}_i\big((\mathbf{X}_i^2)'(\mathbf{y}_i - \mathbf{X}_i^1\boldsymbol{\alpha})/\sigma_\varepsilon^2 + \mathbf{Q}_i^{-1}\boldsymbol{\beta}_{S_i}^G\big)$$

and

$$\mathbf{B}_i = \big((\mathbf{X}_i^2)'\mathbf{X}_i^2/\sigma_\varepsilon^2 + \mathbf{Q}_i^{-1}\big)^{-1},$$

where $\mathbf{Q}_i = \mathbf{Q}_{S_i}^G$.

To sample from the posterior $\pi(\boldsymbol{\phi}_L|\mathbf{y}^N, \boldsymbol{\phi}_V, \mathbf{S}^N)$, we reconsider the marginal model from (13) with the random effects integrated out. An alternative way to write this model is

$$\mathbf{y}_i = \mathbf{Z}_i^*\boldsymbol{\alpha}^* + \boldsymbol{\varepsilon}_i^*, \qquad \boldsymbol{\varepsilon}_i^* \sim N(\mathbf{0}, \mathbf{V}_i), \tag{B.1}$$

with $\mathbf{Z}_i^* = (\mathbf{X}_i^1\ \mathbf{X}_i^2D_i^{(1)} \ldots \mathbf{X}_i^2D_i^{(K)})$, where we used the coding $D_i^{(k)} = 1$ iff $S_i = k$, for $k = 1, \ldots, K$. Furthermore, $\mathbf{V}_i = \mathbf{X}_i^2\mathbf{Q}_{S_i}^G(\mathbf{X}_i^2)' + \sigma_\varepsilon^2\mathbf{I}$.

From the marginal model (B.1), we see that the posterior of $\boldsymbol{\phi}_L$ is normally distributed with $\pi(\boldsymbol{\phi}_L|\mathbf{y}^N, \boldsymbol{\phi}_V, \mathbf{S}^N) \sim N(\mathbf{a}_N^*, \mathbf{A}_N^*)$, where

$$\mathbf{A}_N^* = \left(\sum_{i=1}^N (\mathbf{Z}_i^*)'\mathbf{V}_i^{-1}\mathbf{Z}_i^* + (\mathbf{A}_0^*)^{-1}\right)^{-1} \tag{B.2}$$

and

$$\mathbf{a}_N^* = \mathbf{A}_N^*\left(\sum_{i=1}^N (\mathbf{Z}_i^*)'\mathbf{V}_i^{-1}\mathbf{y}_i + (\mathbf{A}_0^*)^{-1}\mathbf{a}_0^*\right), \tag{B.3}$$

with $\mathbf{a}_0^* = (\mathbf{a}_0'\ \underbrace{\mathbf{c}_0' \ldots \mathbf{c}_0'}_{K})'$ and $\mathbf{A}_0^*$ having $\mathbf{A}_0$ and K blocks of $\mathbf{C}_0$ on its diagonal,

$$\mathbf{A}_0^* = \begin{pmatrix}\mathbf{A}_0 & 0 & 0 & 0\\ 0 & \mathbf{C}_0 & 0 & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \mathbf{C}_0\end{pmatrix}.$$

Here $\boldsymbol{\phi}_L$ may be sampled directly from (B.2) and (B.3). The information matrix $(\mathbf{A}_N^*)^{-1}$, however, has a special structure that can be exploited for efficient sampling. If no fixed effects $\boldsymbol{\alpha}$ are present, then $(\mathbf{A}_N^*)^{-1}$ and $\mathbf{A}_N^*$ are block diagonal, and we may sample all group-specific effects independently. If fixed effects $\boldsymbol{\alpha}$ are present, then $(\mathbf{A}_N^*)^{-1}$ and $\mathbf{A}_N^*$ contain a submatrix that is block diagonal. Therefore, joint sampling of $\boldsymbol{\phi}_L = (\boldsymbol{\alpha}, \boldsymbol{\beta}_1^G, \ldots, \boldsymbol{\beta}_K^G)$ is possible by sampling the fixed effects from the marginal posterior $N(\mathbf{c}_N, \mathbf{C}_N)$, where the group-specific effects are integrated out, and by sampling the group-specific effects independently from the conditional distributions $N(\mathbf{b}_{N,k}(\boldsymbol{\alpha}), \mathbf{B}_{N,k}), k = 1, \ldots, K$. The moments of these densities are given by

$$\mathbf{b}_{N,k}(\boldsymbol{\alpha}) = \mathbf{B}_{N,k}\left[\sum_{i=1}^N (\mathbf{X}_i^2)'\mathbf{V}_i^{-1}D_i^{(k)}\tilde{\mathbf{y}}_i + \mathbf{B}_0^{-1}\mathbf{b}_0\right],$$

$$\mathbf{B}_{N,k} = \left[\sum_{i=1}^N (\mathbf{X}_i^2)'\mathbf{V}_i^{-1}\mathbf{X}_i^2D_i^{(k)} + \mathbf{B}_0^{-1}\right]^{-1},$$

$$\mathbf{c}_N = \mathbf{C}_N\left[\sum_{i=1}^N (\mathbf{X}_i^1)'\mathbf{V}_i^{-1}\big(\mathbf{y}_i - \mathbf{X}_i^2\mathbf{b}_{N,S_i}(\mathbf{0})\big) + \mathbf{C}_0^{-1}\mathbf{c}_0\right],$$

and

$$\mathbf{C}_N = \left[\sum_{i=1}^N (\mathbf{X}_i^1)'\mathbf{V}_i^{-1}\big(\mathbf{V}_i - \mathbf{X}_i^2\mathbf{B}_{N,S_i}(\mathbf{X}_i^2)'\big)\mathbf{V}_i^{-1}\mathbf{X}_i^1 + \mathbf{C}_0^{-1}\right]^{-1},$$

where $\tilde{\mathbf{y}}_i = \mathbf{y}_i - \mathbf{Z}_i\boldsymbol{\alpha}$.

B.2 Sampling $\mathbf{S}^N$

Applying Bayes' theorem, we obtain

$$\pi(\mathbf{S}^N|\mathbf{y}^N, \boldsymbol{\phi}) \propto \prod_{i=1}^N L\big(\mathbf{y}_i|S_i, \boldsymbol{\alpha}, \boldsymbol{\beta}_{S_i}^G, \mathbf{Q}_{S_i}^G, \sigma_\varepsilon^2\big)\pi(S_i|\boldsymbol{\eta}).$$

Therefore, $S_1, \ldots, S_N$ are conditionally independent given $\boldsymbol{\phi}$ and $\mathbf{y}^N$, and we sample S_i for $i = 1, \ldots, N$ from the discrete distribution

$$\pi(S_i = k|\mathbf{y}_i, \boldsymbol{\phi}) \propto L(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}_k^G, \mathbf{Q}_k^G, \sigma_\varepsilon^2) \cdot \eta_k. \tag{B.4}$$

To obtain the likelihood, we use the marginal model (13) where the random effects are integrated out. Therefore, the likelihood in (B.4) is the likelihood of the normal distribution $N(\mathbf{y}_i; \mathbf{X}_i^1\boldsymbol{\alpha} + \mathbf{X}_i^2\boldsymbol{\beta}_k^G, \mathbf{X}_i^2\mathbf{Q}_k^G(\mathbf{X}_i^2)' + \sigma_\varepsilon^2\mathbf{I})$.

B.3 Sampling the Variance Parameters

Although sampling the variance parameters $\boldsymbol{\phi}_V = (\mathbf{Q}_1^G, \ldots, \mathbf{Q}_K^G, \sigma_\varepsilon^2)$ conditional on $\boldsymbol{\beta}^N$, $\boldsymbol{\phi}_L$, $\mathbf{S}^N$, and $\mathbf{y}^N$ is identical to the corresponding step in Algorithm 1 applied by Lenk and DeSarbo (2000), we give details here as a matter of reference. Conditional on $\boldsymbol{\beta}^N$, $\boldsymbol{\phi}_L$, $\mathbf{S}^N$, and $\mathbf{y}^N$, the parameters $\mathbf{Q}_1^G, \ldots, \mathbf{Q}_K^G, \sigma_\varepsilon^2$ are pairwise independent. The conditional posterior of $\mathbf{Q}_k^G$ given $\boldsymbol{\beta}^N$, $\boldsymbol{\phi}_L$, $\mathbf{S}^N$, and $\mathbf{y}^N$ is an inverted Wishart

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015

density being independent of $\mathbf{y}^N$ and all group-specific parameters but $\boldsymbol{\beta}_k^G$:

$$\mathbf{Q}_k^G|\boldsymbol{\beta}_k^G, \mathbf{S}^N, \boldsymbol{\beta}^N \sim IW(\nu_k^Q, \mathbf{S}_k^Q), \tag{B.5}$$

$$\nu_k^Q = \nu_0^Q + \frac{1}{2}\sum_{i=1}^N I_{\{S_i=k\}},$$

$$\mathbf{S}_k^Q = \mathbf{S}_0^Q + \frac{1}{2}\sum_{i=1}^N (\boldsymbol{\beta}_i - \boldsymbol{\beta}_k^G)(\boldsymbol{\beta}_i - \boldsymbol{\beta}_k^G)' I_{\{S_i=k\}}.$$

The conditional posterior of σ_ε^2 given $\boldsymbol{\beta}^N$, $\boldsymbol{\phi}_L$, $\mathbf{S}^N$, and $\mathbf{y}^N$ is the following inverted-gamma density:

$$\sigma_\varepsilon^2|\boldsymbol{\beta}^N, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N \sim IG(\nu_{\varepsilon,N}, S_{\varepsilon,N}), \tag{B.6}$$

$$\nu_{\varepsilon,N} = \nu_{\varepsilon,0} + \frac{1}{2}\sum_{i=1}^N T_i,$$

$$S_{\varepsilon,N} = S_{\varepsilon,0} + \frac{1}{2}\|\mathbf{y}_i - \mathbf{X}_i^1\boldsymbol{\alpha} - \mathbf{X}_i^2\boldsymbol{\beta}_i\|_2^2.$$

APPENDIX C: DETAILS OF ALGORITHM 3

Here we give details on how to sample the variance parameters $(\mathbf{Q}_1^G, \dots, \mathbf{Q}_K^G, \sigma_\varepsilon^2)$ conditional on $\boldsymbol{\phi}_L$, $\mathbf{S}^N$, $\mathbf{y}^N$ within step (iv-3) of Algorithm 3:

(iv-3a) For $k = 1, \dots, K$, sample $\mathbf{Q}_k^G$ from $\pi(\mathbf{Q}_k^G|\mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$, where $\mathbf{Q}_{-k}^G$ denotes all covariances but $\mathbf{Q}_k^G$.

(iv-3b) Sample σ_ε^2 from $\pi(\sigma_\varepsilon^2|\mathbf{Q}_1^G, \dots, \mathbf{Q}_K^G, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$.

To sample $\mathbf{Q}_k^G$ from $\pi(\mathbf{Q}_k^G|\mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$, we use the Metropolis–Hastings algorithm. Let $(\mathbf{Q}_k^G)^{old}$ be the current value of $\mathbf{Q}_k^G$. We sample a new covariance matrix $(\mathbf{Q}_k^G)^{new}$ from a proposal density $q(\mathbf{Q}_k^G|(\mathbf{Q}_k^G)^{old})$ and accept the new value with probability

$$\begin{aligned}\operatorname{acc}\big((\mathbf{Q}_k^G)^{new}|(\mathbf{Q}_k^G)^{old}\big) &= \min\left(1, \frac{\pi(\mathbf{Q}_1^G, \dots, (\mathbf{Q}_k^G)^{new}, \dots, \mathbf{Q}_K^G, \sigma_\varepsilon^2|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)}{\pi(\mathbf{Q}_1^G, \dots, (\mathbf{Q}_k^G)^{old}, \dots, \mathbf{Q}_K^G, \sigma_\varepsilon^2|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)}\right.\\ &\quad \left.\cdot \frac{q((\mathbf{Q}_k^G)^{old}|(\mathbf{Q}_k^G)^{new})}{q((\mathbf{Q}_k^G)^{new}|(\mathbf{Q}_k^G)^{old})}\right).\end{aligned} \tag{C.1}$$

To compute the ratio, we use the facts that the posterior $\pi(\mathbf{Q}_1^G, \dots, \mathbf{Q}_K^G, \sigma_\varepsilon^2|\boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$ is proportional to the quantity in (14) and that the unknown normalizing constant cancels.

The design of a suitable proposal density turned out to be somewhat challenging. We use a mixture of inverted Wishart proposal,

$$q\big(\mathbf{Q}_k^G|(\mathbf{Q}_k^G)^{old}\big) = \frac{1}{J}\sum_{j=1}^J \pi\big(\mathbf{Q}_k^G|(\boldsymbol{\beta}^N)^j, \boldsymbol{\beta}_k^G, \mathbf{S}^N\big), \tag{C.2}$$

where $\pi(\mathbf{Q}_k^G|(\boldsymbol{\beta}^N)^j, \boldsymbol{\beta}_k^G, \mathbf{S}^N)$ is the conditional posterior of $\mathbf{Q}_k^G$ given that the random effects are equal to $(\boldsymbol{\beta}^N)^j$. This posterior is an inverted Wishart density and appeared both in Algorithm 1 and Algorithm 2 [see (B.5)]. $(\boldsymbol{\beta}^N)^j$ in (C.2) is sampled from the conditional posterior $\pi(\boldsymbol{\beta}^N|\boldsymbol{\phi}_L, \mathbf{S}^N, (\mathbf{Q}_k^G)^{old}, \mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \mathbf{y}^N)$. The rationale behind this choice is that as $J \to \infty$, the proposal approaches the conditional marginal posterior $\pi(\mathbf{Q}_k^G|\mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \boldsymbol{\phi}_L, \mathbf{S}^N, \mathbf{y}^N)$ that we require to sample from.

The posterior $\pi(\boldsymbol{\beta}^N|\boldsymbol{\phi}_L, \mathbf{S}^N, (\mathbf{Q}_k^G)^{old}, \mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \mathbf{y}^N)$ is the same as in step (iii-3b) of this algorithm. Therefore, to obtain $(\boldsymbol{\beta}^N)^j$, $j = 1, \dots, J$, we repeat step (iii-3b) J times to construct the proposal density $q(\mathbf{Q}_k^G|(\mathbf{Q}_k^G)^{old})$. To sample $(\mathbf{Q}_k^G)^{new}$ from $q(\mathbf{Q}_k^G|(\mathbf{Q}_k^G)^{old})$, we first randomly select a component $\tilde{j}$, and then sample $(\mathbf{Q}_k^G)^{new}$ from the conditional inverted Wishart density $\pi(\mathbf{Q}_k^G|(\boldsymbol{\beta}^N)^{\tilde{j}}, \boldsymbol{\beta}_k^G, \mathbf{S}^N)$. Note that for evaluating $q((\mathbf{Q}_k^G)^{old}|(\mathbf{Q}_k^G)^{new})$ in the acceptance ratio (C.1), we need to sample further J values $(\boldsymbol{\beta}^N)^j$ from the conditional posterior $\pi(\boldsymbol{\beta}^N|\boldsymbol{\phi}_L, \mathbf{S}^N, (\mathbf{Q}_k^G)^{new}, \mathbf{Q}_{-k}^G, \sigma_\varepsilon^2, \mathbf{y}^N)$, where the covariance matrix of group k is substituted by $(\mathbf{Q}_k^G)^{new}$. Increasing J will increase the acceptance rate of the Metropolis–Hastings algorithm; however, choosing J too large will slow the algorithm, because for each step and each group we need $2J$ functional evaluations of an inverted Wishart density.

A comparable Metropolis–Hastings algorithm is used to sample σ_ε^2.

APPENDIX D: PROOF OF (21)

Equation (21) is easy to verify. First,

$$\begin{aligned}E_{\mathcal{A}}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big) &= \int_{\mathcal{A}} f(\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi})\\ &= \sum_{s=1}^{K!}\int_{\mathcal{L}_s} f(\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi}).\end{aligned}$$

Because both $f(\boldsymbol{\psi})$ and $\pi(\boldsymbol{\psi}|\mathbf{y}^N)$ are invariant to relabeling, we obtain $\int_{\mathcal{L}_s} f(\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi}) = \int_{\mathcal{L}_l} f(\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi})$ for all l and s, and the following holds:

$$E_{\mathcal{A}}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big) = K!\int_{\mathcal{L}_l} f(\boldsymbol{\psi})\pi(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi}). \tag{D.1}$$

If we take the special case where $f(\boldsymbol{\psi}) = 1$, then we obtain

$$1 = K!\int_{\mathcal{L}_l} \pi(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi}).$$

Therefore, the posterior distribution $\pi_{\mathcal{L}_l}(\boldsymbol{\psi}|\mathbf{y}^N)$, which is constrained to the labeling subspace $\mathcal{L}_l$, is related to the unconstrained posterior $\pi(\boldsymbol{\psi}|\mathbf{y}^N)$ by $\pi_{\mathcal{L}_l}(\boldsymbol{\psi}|\mathbf{y}^N) = K!\pi(\boldsymbol{\psi}|\mathbf{y}^N)$. Therefore, (21) follows from (D.1),

$$\begin{aligned}E_{\mathcal{A}}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big) &= \int_{\mathcal{L}_l} f(\boldsymbol{\psi})\pi_{\mathcal{L}_l}(\boldsymbol{\psi}|\mathbf{y}^N)\nu(d\boldsymbol{\psi})\\ &= E_{\mathcal{L}_s}\big(f(\boldsymbol{\psi})|\mathbf{y}^N\big).\end{aligned}$$

[Received May 2000. Revised May 2003.]

REFERENCES

Aitkin, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 47, 67–75.

Allenby, G. M., Arora, N., and Ginter, J. L. (1998), "On the Heterogeneity of Demand," *Journal of Marketing Research*, 35, 384–389.

Celeux, G., Hurn, M., and Robert, C. P. (2000), "Computational and Inferential Difficulties with Mixture Posterior Distributions," *Journal of the American Statistical Association*, 95, 957–970.

Chen, M., Shao, Q., and Ibrahim, J. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer.

Chib, S., and Carlin, B. (1999), "On MCMC Sampling in Hierarchical Longitudinal Models," *Statistics and Computing*, 9, 17–26.

Elrod, T., Louviere, J. J., and Davey, K. S. (1992), "An Empirical Comparison of Ratings-Based and Choice-Based Conjoint Models," *Journal of Marketing Research*, 29, 368–377.

Frühwirth-Schnatter, S. (1999), "Model Likelihoods and Bayes Factors for Switching and Mixture Models," preprint, Vienna University of Economics and Business Administration, revised version submitted for publication.

——— (2001a), "Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models," *Journal of the American Statistical Association,* 96, 194–209.

——— (2001b), "Fully Bayesian Analysis of Switching Gaussian State-Space Models," *Annals of the Institute of Mathematical Statistics*, 53, 31–49.

Frühwirth-Schnatter, S., and Otter, T. (1999), "Conjoint Analysis Using Mixed Effect Models," in *Statistical Modelling, Proceedings of the Fourteenth International Workshop on Statistical Modelling*, eds. H. Friedl, A. Berghold, and G. Kauermann, Graz, pp. 181–191.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parameterizations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

Gerlach, R., Carter, C., and Kohn, R. (2000), "Efficient Bayesian Inference for Dynamic Mixture Models," *Journal of the American Statistical Association*, 95, 819–828.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in *Bayesian Statistics* (Vol. 4), eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 169–193.

Kaufmann, S., and Frühwirth-Schnatter, S. (2002), "Bayesian Analysis of Switching ARCH Models," *Journal of Time Series Analysis*, 23, 425–458.

Lenk, P. J., and DeSarbo, W. S. (2000), "Bayesian Inference for Finite Mixture of Generalized Linear Models With Random Effects," *Psychometrika*, 65, 93–119.

Liu, J., Wong, W., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

Meng, X., and Van Dyck, D. (1998), "Fast EM-Type Implementations for Mixed Effects Models," *Journal of the Royal Statistical Society*, Ser. B, 60, 559–578.

Otter, T., Tüchler, R., and Frühwirth-Schnatter, S. (2002), "Bayesian Latent Class Metric Conjoint Analysis—A Case Study from the Austrian Mineral Water Market," in *Explanatory Data Analysis in Empirical Research*, eds. O. Opitz and M. Schwaiger, Berlin: Springer-Verlag, pp. 157–169.

Stephens, M. (2000), "Dealing With Label Switching in Mixture Models," *Journal of the Royal Statistical Society*, Ser. B, 62, 795–809.

Tüchler, R., Frühwirth-Schnatter, S., and Otter, T. (2002), "The Heterogeneity Model and Its Special Cases—An Illustrative Comparison," in *Proceedings of the 17th International Workshop on Statistical Modelling*, eds. M. Stasinopoulos and G. Touloumi, Chania, Greece, pp. 637–644.

Van Dyck, D., and Meng, X. (2001), "The Art of Data Augmentation," *Journal of Computational and Graphical Statistics,* 110, 1–50.

Verbeke, G., and Lesaffre, E. (1996), "A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population," *Journal of the American Statistical Association*, 91, 217–221.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

Downloaded by [KU Leuven University Library] at 02:28 13 October 2015