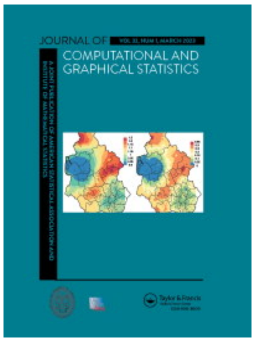

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Streamlined Variational Inference for Linear Mixed Models with Crossed Random Effects

Marianne Menictas, Gioia Di Credico & Matt P. Wand

To cite this article: Marianne Menictas, Gioia Di Credico & Matt P. Wand (2023) Streamlined Variational Inference for Linear Mixed Models with Crossed Random Effects, Journal of Computational and Graphical Statistics, 32:1, 99-115, DOI: 10.1080/10618600.2022.2096622

To link to this article: https://doi.org/10.1080/10618600.2022.2096622

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 32, NO. 1, 99–115
https://doi.org/10.1080/10618600.2022.2096622

Taylor & Francis
Taylor & Francis Group

Check for updates

Streamlined Variational Inference for Linear Mixed Models with Crossed Random Effects

Marianne Menictas[a], Gioia Di Credico[b], and Matt P. Wand[c]

[a]Harvard University, , Cambridge, MA; [b]University of Trieste, Trieste, Italy; [c]University of Technology Sydney, Sydney, Australia

ABSTRACT
We derive streamlined mean field variational Bayes algorithms for fitting linear mixed models with crossed random effects. In the most general situation, where the dimensions of the crossed groups are arbitrarily large, streamlining is hindered by lack of sparseness in the underlying least squares system. Because of this fact we also consider a hierarchy of relaxations of the mean field product restriction. The least stringent product restriction delivers a high degree of inferential accuracy. However, this accuracy must be mitigated against its higher storage and computing demands. Faster sparse storage and computing alternatives are also provided, but come with the price of diminished inferential accuracy. This article provides full algorithmic details of three variational inference strategies, presents detailed empirical results on their pros and cons and, thus, guides the users on their choice of variational inference approach depending on the problem size and computing resources. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2021
Accepted June 2022

KEYWORDS
Item response theory; Mean field variational Bayes; Rasch analysis; Scalable statistical methodology; Sparse least squares systems

1. Introduction

Linear mixed models with crossed random effects are a useful vehicle for analysis and inference for data that are cross-classified according to two or more grouping mechanisms. One major application area is psychometrics in which a cohort of *subjects* is assessed according to a set of tasks or *items* (e.g., Baayen, Davidson, and Bates 2008; Jeon, Rijmen, and Rabe-Hesketh 2017). The assessment scores are cross-classified according to subject and item. In such studies it is common for both the subjects and items to be treated as random samples from relevant populations. For example, in a psycholinguistic study, the subjects may be a random sample from the population of native Greek speakers and the items may be a random sample from the population of Greek language syllables. Other variables such as gender and stimuli type may be treated as nonrandom. Mixed models with crossed random effects for subject and item and fixed effects for variables of interest facilitate inference for Greek speakers and the Greek language in general rather than for the participants and syllables chosen for the study. Other areas of psychometrics such as item response theory and Rasch analysis (e.g., Doran et al. 2007) benefit from crossed random effects models. The essence of this contribution is streamlined variational inference for crossed random effects mixed models that scales well to the handling of very large datasets.

The term "streamlined" refers to the process of taking advantage of sparse structures within the design matrices that arise in linear mixed models. The design matrices are often very sparse and potentially extremely large. Clever algorithms that recognize and make use of the sparseness patterns can lead to dramatic savings in terms of storage and computing time. Nolan et al. (2020) provides a systematic treatment of streamlined variational inference for linear mixed models with two and three levels of nesting. The group specific curves extension is dealt with in Menictas et al. (2021). In these articles, each involving the first and third authors of the current article, it was recognized that key variational inference updates can be embedded with the class of two-level sparse least squares problems (Nolan and Wand 2020) and that this algorithmic component can be isolated into a procedure that we call SolveTwoLevelSparseLeastSquares. This procedure also arises in our variational inference algorithms for crossed random effects in Section 4. Also, Nolan et al. (2020) and Menictas et al. (2021) are concerned with nested random effects models whilst this article treats the crossed random effects situation. The former situation is less challenging since higher level nesting invokes hierarchical sparsity structures that are amenable to streamlined fitting strategies. These strategies are fully efficient in terms of only using the nonzero entries of the design matrices. For crossed random effects the sparsity structure, if present, is more delicate. Depending on the restrictiveness of the variational approach and the cross-tabulation variable sizes, the cross random effects sparseness structure may not be amenable to fully efficient fitting and inference.

Throughout this article we consider two grouping mechanisms with group dimensions m and m'. Furthermore, we label the groups in such a way that $m \geq m'$. For example, a psycholinguistic study involving 900 subjects and 40 items has group sizes $m = 900$ and $m' = 40$. If a different study involved 75 subjects and 80 items then the (m, m') labeling is reversed

CONTACT Marianne Menictas mmenictas@gmail.com Harvard University, Department of Statistics, Cambridge, MA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

with respect to subjects and items and our notation is $m = 80$ items and $m' = 75$ subjects. Sticking with the $m \geq m'$ notation is important, since it affects variational inference algorithm construction and choice. For example, if m' is moderate in size and m is very large then the least squares system that underlies the least stringent (most accurate) variational inference scheme is sparse, and streamlined computing advantages are available. On the other hand, if m' is also very large then the least stringent algorithm is non-sparse and, depending on computing resources and run-time demands, more stringent (less accurate) variational inference schemes may be preferred.

The variational Bayesian inference paradigm is becoming quite a powerful one in contemporary statistical and machine learning contexts (e.g., Blei et al. 2017). Modularization variants such as variational message passing (Winn and Bishop 2005; Wand 2017) have allowed for the development of versatile and fast inference engines such as Edward (Tran et al. 2016) and Infer.NET (Minka et al. 2018). Various options concerning the stringency of mean field-type product restrictions allow for scalability to very large problems with speed being traded off against accuracy. All algorithms presented here are purely matrix algebraic and require no root-finding or numerical integration. Our variational inference algorithm with medium product restrictions is able to handle hundreds of crossed random effects in tens of seconds on contemporary laptop computers.

The use of variational approximations for crossed random effects mixed models is an emerging activity and, to date, there are only a few contributions of this type. The most prominent such contribution is Jeon, Rijmen, and Rabe-Hesketh (2017) which applied the notions of Gaussian variational approximation to frequentist generalized linear mixed models with crossed random effects. Jeon, Rijmen, and Rabe-Hesketh (2017) concentrated on the scalar effects case and also imposed a product restriction between the "item" and "subject" random effects. Our algorithms, which are for approximate Bayesian inference, allow for this restriction to be removed albeit at the cost of increased storage and computation. We also focus on the Gaussian response here and give a thorough treatment of this more straightforward case. Semiparametric mean field variational Bayes ideas (e.g., Nolan and Wand 2017) facilitate extension to other likelihoods.

In Section 2 we define a general class of Gaussian response Bayesian crossed random effects linear mixed models. Sections 3 and 4 form the centerpiece of the article and explain various mean field variational Bayes strategies, followed by listings of algorithms that facilitate streamlined implementation. In Section 5 we report on the results of simulation-based numerical studies that assess and compare the performances of these new algorithms with respect to inferential accuracy and computing time. Section 6 contains an illustration for data from a large longitudinal education study. We summarize our findings in Section 7. The supplementary materials contains derivational and related details. Some results for frequentist inference for crossed random effects are also given in the supplementary materials.

2. Bayesian Crossed Random Effects Linear Mixed Models

The Bayesian crossed random effects linear mixed models being considered here are such that:

$$\begin{aligned}\boldsymbol{y}_{ii'}|\boldsymbol{\beta},\boldsymbol{u}_i,\boldsymbol{u}'_{i'},\sigma^2 &\overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{ii'}\boldsymbol{\beta}+\boldsymbol{Z}_{ii'}\boldsymbol{u}_i+\boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'},\sigma^2\boldsymbol{I}),\\ \boldsymbol{u}_i|\boldsymbol{\Sigma} \overset{\text{ind.}}{\sim} N(\boldsymbol{0},\boldsymbol{\Sigma}),\quad 1\le i\le m,&\quad \boldsymbol{u}'_{i'}|\boldsymbol{\Sigma}'\overset{\text{ind.}}{\sim} N(\boldsymbol{0},\boldsymbol{\Sigma}'),\\ 1\le i'\le m',&\quad \boldsymbol{\beta}\sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}}).\end{aligned} \tag{1}$$

The matrices in (1) have dimensions as follows:

$$\begin{gathered}\boldsymbol{y}_{ii'} \text{ is } n_{ii'}\times 1,\ \ \boldsymbol{X}_{ii'} \text{ is } n_{ii'}\times p,\ \ \boldsymbol{\beta} \text{ is } p\times 1,\\ \boldsymbol{Z}_{ii'} \text{ is } n_{ii'}\times q,\ \ \boldsymbol{u}_i \text{ is } q\times 1,\ \ \boldsymbol{Z}'_{ii'} \text{ is } n_{ii'}\times q',\ \ \boldsymbol{u}'_{i'} \text{ is } q'\times 1,\\ \boldsymbol{\Sigma} \text{ is } q\times q \text{ and } \boldsymbol{\Sigma}' \text{ is } q'\times q'.\end{gathered} \tag{2}$$

Here $n_{ii'}$ is the number of response measurements in the (i, i')th cell. If $n_{ii'} = 0$ then each of $\boldsymbol{y}_{ii'}$, $\boldsymbol{X}_{ii'}$, $\boldsymbol{Z}_{ii'}$ and $\boldsymbol{Z}'_{ii'}$ are null. However, for upcoming matrix assembly operations it is useful to think of, $\boldsymbol{Z}_{ii'}$, for example, as an $n_{ii'} \times q$ "matrix" with $n_{ii'} = 0$.

To aid digestibility of (1) and (2), consider a generic education research study where a sample of m students is followed longitudinally and have academic performances measured according to m' items, such as those which quantify cognitive, literary and numeracy abilities. The items take the form of exercises and, for each item, a quantitative score is determined from a student's performance in that item's exercises. Over the duration of the multi-year study each of the m students are scored on the m' items n times, which implies that $n_{ii'} = n$ for all $1 \le i \le m$ and $1 \le i' \le m'$. Define $\texttt{score}_{ii'j}$ to be the jth score of student i for item i'. Let $\texttt{age}_{ii'j}$ be defined analogously, corresponding to age in years. Lastly, define $\texttt{train}_{ii'j}$ to be the indicator of whether the ith student received training prior to their jth attempt at the i'th item. Then a $p = 3$ and $q = q' = 2$ version of the response vector and design matrices is

$$\boldsymbol{y}_{ii'} = \left[\texttt{score}_{ii'j}\right]_{1\le j\le n},\quad \boldsymbol{X}_{ii'} = \left[1\ \texttt{age}_{ii'j}\ \texttt{train}_{ii'j}\right]_{1\le j\le n},$$
$$\boldsymbol{Z}_{ii'} = \left[1\ \texttt{age}_{ii'j}\right]_{1\le j\le n}$$

with $\boldsymbol{Z}'_{ii'} = \boldsymbol{Z}_{ii'}$. According to (1) and this set-up, the scores of the ith student on the i'th item are modeled to be

$$\begin{aligned} &y_{ii'j}|\beta_0,\beta_1,\beta_2,u_{0i},u_{1i},u'_{0i'},u'_{1i'},\sigma^2\\ &\quad\overset{\text{ind.}}{\sim} N\big((\beta_0+u_{0i}+u'_{0i'})+(\beta_1+u_{1i}+u'_{1i'})\texttt{age}_{ii'j}\\ &\quad+\beta_2\texttt{train}_{ii'j},\sigma^2\big),\quad 1\le j\le n.\end{aligned}$$

Conditional on $\boldsymbol{\Sigma}$, the $[u_{0i}\ u_{1i}]^T$ are $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ random vectors. The $[u'_{0i'}\ u'_{1i'}]^T$ are similar with $\boldsymbol{\Sigma}'$ instead of $\boldsymbol{\Sigma}$. It is apparent from this that model (1) allows for a different intercept and slope for every subject/item combination. The heterogeneities in the intercepts and slopes correspond to appropriate entries of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$. If the fixed effect β_2 is of primary interest then (1) is a parsimonious model that allows for subject/item heterogeneities in the age effects.

For the error variance σ^2 and the random effects covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ we consider two prior distribution families:

(A) ordinary Inverse-Wishart priors

(B) the marginally non-informative priors proposed in Huang and Wand (2013).

In terms of the Inverse Chi-squared and Inverse-G-Wishart distributional notation given in Section S.1 of the supplementary materials, prior specification (A) involves:

$$\begin{aligned}&\sigma^2 \sim \text{Inverse-}\chi^2(\xi_{\sigma^2},\lambda_{\sigma^2}),\\ &\boldsymbol{\Sigma} \sim \text{Inverse-G-Wishart}(G_{\text{full}},\xi_{\boldsymbol{\Sigma}},\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}),\\ &\boldsymbol{\Sigma}' \sim \text{Inverse-G-Wishart}(G_{\text{full}},\xi_{\boldsymbol{\Sigma}'},\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}'})\end{aligned} \tag{3}$$

for hyperparameters $\xi_{\sigma^2}, \lambda_{\sigma^2} > 0, \xi_{\boldsymbol{\Sigma}} > 2(q-1), \xi_{\boldsymbol{\Sigma}'} > 2(q'-1)$ and symmetric positive definite matrices $\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}'}$. Prior specification (B) involves:

$$
\begin{aligned}
&\sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(\nu_{\sigma^2}, 1/a_{\sigma^2}),\\
&a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\sigma^2}s^2_{\sigma^2})),\\
&\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}} + 2q - 2, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}),\\
&\boldsymbol{\Sigma}'|\boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}'} + 2q' - 2, \boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1}),\\
&\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{A_{\boldsymbol{\Sigma}}}),\\
&\boldsymbol{\Lambda}_{A_{\boldsymbol{\Sigma}}} \equiv \left\{\nu_{\boldsymbol{\Sigma}}\text{diag}(s^2_{\boldsymbol{\Sigma},1}, \ldots, s^2_{\boldsymbol{\Sigma},q})\right\}^{-1}\\
&\boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{A_{\boldsymbol{\Sigma}'}}),\\
&\boldsymbol{\Lambda}_{A_{\boldsymbol{\Sigma}'}} \equiv \left\{\nu_{\boldsymbol{\Sigma}'}\text{diag}(s^2_{\boldsymbol{\Sigma}',1}, \ldots, s^2_{\boldsymbol{\Sigma}',q'})\right\}^{-1}
\end{aligned}
\tag{4}
$$

for hyperparameters $\nu_{\sigma^2}, \nu_{\boldsymbol{\Sigma}}, \nu_{\boldsymbol{\Sigma}'}, s^2_{\boldsymbol{\Sigma},1}, \ldots, s^2_{\boldsymbol{\Sigma},q}, s^2_{\boldsymbol{\Sigma}',1}, \ldots, s^2_{\boldsymbol{\Sigma}',q'} > 0$. As explained in Huang and Wand (2013), such priors allow standard deviation and correlation parameters to have arbitrary non-informativeness.

2.1. Additional Data Matrices

The various streamlined mean field variational Bayes algorithms given in Section 4 benefit from the setting up of additional data matrices in which the raw data in $\boldsymbol{y}_{ii'}$, $\boldsymbol{X}_{ii'}$, $\boldsymbol{Z}_{ii'}$ and $\boldsymbol{Z}'_{ii'}$ are combined in various ways using "stack" and "blockdiag" operators. These operators are defined as follows:

$$
\underset{1\le i\le d}{\text{stack}}(\boldsymbol{M}_i) \equiv \begin{bmatrix} \boldsymbol{M}_1 \\ \vdots \\ \boldsymbol{M}_d \end{bmatrix} \quad \text{and}
$$

$$
\underset{1\le i\le d}{\text{blockdiag}}(\boldsymbol{M}_i) \equiv \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{M}_d \end{bmatrix}
$$

for matrices $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_d$. The first of these definitions require that $\boldsymbol{M}_i$, $1 \le i \le d$, each have the same number of columns. For the null design matrices that may arise in crossed random effects models it is convenient to adopt generalizations of regular matrix manipulations. If one of the $\boldsymbol{M}_i$ is $n \times p$ where $n = 0$ and $p > 0$ then it is ignored by the stack operator. However, for the blockdiag operator the column index should have an increment of p before adding the next matrix. This subtledy is fully explained in Section S.2 of the supplementary materials. To appreciate the motivation for the "stack" and "blockdiag" notation, consider the intercepts-only special case where $p = q = q' = 1$, $m = m' = 2$ and $n_{ii'} = 1$ for $i, i' = 1, 2$. Then the full set of conditional means is contained in the vector

$$
\begin{bmatrix} \beta_0 + u_{01} + u'_{01} \\ \beta_0 + u_{01} + u'_{02} \\ \beta_0 + u_{02} + u'_{01} \\ \beta_0 + u_{02} + u'_{02} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ u_{01} \\ u_{02} \\ u'_{01} \\ u'_{02} \end{bmatrix}. \tag{5}
$$

Note that the design matrix in (5) can be written as

$$
\left[\mathbf{1}_4 \;\; \underset{1\le i\le 2}{\text{blockdiag}}\, \mathbf{1}_2 \;\; \underset{1\le i\le 2}{\text{stack}}\, \boldsymbol{I}_2\right].
$$

where $\mathbf{1}_d$ denotes the $d \times 1$ vector of ones and $\boldsymbol{I}_d$ is the $d \times d$ identity matrix. It is apparent from this example that such notation is very useful for handling cross random effects design structures. The remainder of this section allows for similar organization of the response and predictor data and greatly aids succinct algorithmic description, which involve various full conditional distributions.

Our first set of additional data matrices is

$$
\overset{\blacktriangle}{\boldsymbol{y}}_i \equiv \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{y}_{ii'}), \quad \overset{\blacktriangle}{\boldsymbol{X}}_i \equiv \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{X}_{ii'}), \quad 1 \le i \le m,
$$

and

$$
\overset{\blacktriangledown}{\boldsymbol{y}}_{i'} \equiv \underset{1\le i\le m}{\text{stack}}(\boldsymbol{y}_{ii'}), \quad \overset{\blacktriangledown}{\boldsymbol{X}}_{i'} \equiv \underset{1\le i\le m}{\text{stack}}(\boldsymbol{X}_{ii'}), \quad 1 \le i' \le m'.
$$

Next define

$$
\overset{\blacktriangle}{\boldsymbol{Z}}_i \equiv \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{Z}_{ii'}), \overset{\blacksquare}{\boldsymbol{Z}'_i} \equiv \underset{1\le i'\le m'}{\text{blockdiag}}(\boldsymbol{Z}'_{ii'}), \; 1 \le i \le m, \quad \text{and}
$$

$$
\overset{\blacktriangledown}{\boldsymbol{Z}'}_{i'} \equiv \underset{1\le i\le m}{\text{stack}}(\boldsymbol{Z}'_{ii'}), \; 1 \le i' \le m'.
$$

Also, we define

$$
\begin{aligned}
\boldsymbol{y} &\equiv \underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{y}_{ii'})\right\} = \underset{1\le i\le m}{\text{stack}}(\overset{\blacktriangle}{\boldsymbol{y}}_i),\\
\boldsymbol{X} &\equiv \underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{X}_{ii'})\right\} = \underset{1\le i\le m}{\text{stack}}(\overset{\blacktriangle}{\boldsymbol{X}}_i)
\end{aligned}
$$

and

$$
\boldsymbol{Z} \equiv \left[\text{blockdiag}_{1\le i\le m}(\overset{\blacktriangle}{\boldsymbol{Z}}_i) \;\; \text{stack}_{1\le i\le m}(\overset{\blacksquare}{\boldsymbol{Z}'_i})\right].
$$

2.2. Additional Dimensional Notation

The dimensions of the data matrices defined in Section 2.1 are such that the following notation is useful:

$$
\begin{aligned}
n_{i\bullet} &\equiv \sum_{i'=1}^{m'} n_{ii'}, \quad 1 \le i \le m,\\
n_{\bullet i'} &\equiv \sum_{i=1}^{m} n_{ii'}, \quad 1 \le i' \le m', \quad \text{and}\\
n_{\bullet\bullet} &\equiv \sum_{i=1}^{m}\sum_{i'=1}^{m'} n_{ii'}.
\end{aligned}
$$

3. Variational Inference

The joint conditional density function of all parameters in (1) with covariance priors (3) is

$$
\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'|\boldsymbol{y}). \tag{6}
$$

where $\boldsymbol{u} \equiv (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ and $\boldsymbol{u}' \equiv (\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_{m'})$. Let

$$
\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}') \tag{7}
$$

be a mean field approximation of (6). Several product restrictions can be placed on the q-density function in (7). Here we consider three such restrictions:

$$\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}') = \begin{cases} \mathrm{q}(\boldsymbol{\beta})\mathrm{q}(\boldsymbol{u})\mathrm{q}(\boldsymbol{u}')\,\mathrm{q}(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'), & \text{labeled } \textit{product restriction I,}\\ \mathrm{q}(\boldsymbol{\beta},\boldsymbol{u})\mathrm{q}(\boldsymbol{u}')\,\mathrm{q}(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'), & \text{labeled } \textit{product restriction II,}\\ \mathrm{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')\,\mathrm{q}(\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'), & \text{labeled } \textit{product restriction III.}\end{cases} \quad (8)$$

Product restriction I has the simplest streamlined implementation but it sets all posterior correlations between $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{u}'$ to zero and, thus, produces posterior distributions with overly small variances. On the other hand, product restriction III allows for joint posterior covariance matrix of $(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')$ in its q-density to be full—which leads to higher inferential accuracy but more challenging computing that can only be streamlined if m' is moderate. Product restriction II is a halfway house that recognizes the $m \geq m'$ asymmetry and carries posterior correlations between $\boldsymbol{\beta}$ and $\boldsymbol{u}$, which is the larger of $\boldsymbol{u}$ and $\boldsymbol{u}'$ assuming that q and q' have similar sizes. It delivers more accurate inference than product restriction I but with similar computational overhead.

It should be noted that (8) conveys the product restrictions in their minimal forms. However, conditional independencies inherent in (1) mean that additional factorizations ensue as follows:

$$\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}') = \begin{cases} \mathrm{q}(\boldsymbol{\beta})\left\{\prod_{i=1}^{m}\mathrm{q}(\boldsymbol{u}_i)\right\}\left\{\prod_{i'=1}^{m'}\mathrm{q}(\boldsymbol{u}'_{i'})\right\} \\ \quad\times\mathrm{q}(\sigma^2)\mathrm{q}(\boldsymbol{\Sigma})\mathrm{q}(\boldsymbol{\Sigma}'), & \text{for product restriction I,}\\ \mathrm{q}(\boldsymbol{\beta},\boldsymbol{u})\left\{\prod_{i'=1}^{m'}\mathrm{q}(\boldsymbol{u}'_{i'})\right\}\mathrm{q}(\sigma^2)\mathrm{q}(\boldsymbol{\Sigma})\mathrm{q}(\boldsymbol{\Sigma}'), & \text{for product restriction II,}\\ \mathrm{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')\mathrm{q}(\sigma^2)\mathrm{q}(\boldsymbol{\Sigma})\mathrm{q}(\boldsymbol{\Sigma}'), & \text{for product restriction III.}\end{cases}$$

If, instead, the Huang and Wand (2013) priors are used then conditional independencies inherent in (4) lead to the covariance matrix and auxiliary variables component of the joint q-density factorizing fully as follows:

$$\mathrm{q}(\sigma^2,a_{\sigma^2},\boldsymbol{\Sigma},\boldsymbol{A_\Sigma},\boldsymbol{\Sigma}',\boldsymbol{A_{\Sigma'}}) = \mathrm{q}(\sigma^2)\mathrm{q}(a_{\sigma^2})\mathrm{q}(\boldsymbol{\Sigma})\mathrm{q}(\boldsymbol{A_\Sigma})\mathrm{q}(\boldsymbol{\Sigma}')\mathrm{q}(\boldsymbol{A_{\Sigma'}}).$$

Under either product restrictions I, II, or III, and letting $\boldsymbol{u}_{\rm all} \equiv (\boldsymbol{u},\boldsymbol{u}')$, standard mean field variational Bayes steps (e.g., Bishop 2006, sec. 10.1–10.3) lead to the q-density functions of the model parameters having the following forms:

$\mathrm{q}^*(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})$ has a $N\big(\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})},\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}\big)$ distribution,
$\mathrm{q}^*(\sigma^2)$ has an Inverse-$\chi^2\big(\xi_{\mathrm{q}(\sigma^2)},\lambda_{\mathrm{q}(\sigma^2)}\big)$ distribution,
$\mathrm{q}^*(\boldsymbol{\Sigma})$ has an Inverse-G-Wishart$(G_{\rm full},\xi_{\mathrm{q}(\boldsymbol{\Sigma})},\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})})$ distribution
and $\mathrm{q}^*(\boldsymbol{\Sigma}')$ has an Inverse-G-Wishart$(G_{\rm full},\xi_{\mathrm{q}(\boldsymbol{\Sigma}')},\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')})$ distribution.

The q-density parameters can be obtained using a coordinate ascent iterative algorithm (e.g., Algorithm 1 of Ormerod and Wand (2010). However, if applied naïvely, the matrix $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ requires storage and inversion. As explained in the upcoming Section 3.1, this matrix is potentially prohibitively large. Product restrictions I, II, and III lead to streamlined mean field variational Bayes algorithms with varying degrees of storage and computational overhead.

Table 1. The zero ($\boldsymbol{O}$) versus nonzero ($\times$) status of various sub-blocks of the $\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})$ covariance matrix $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ under product restrictions I, II, and III.

sub-blocks of $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$	prod. res. I	prod. res. II	prod. res. III
$\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})}$, $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)}$, $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}'_{i'})}$	$\times$	$\times$	$\times$
$E_{\mathrm{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T\}$	$\boldsymbol{O}$	$\times$	$\times$
$E_{\mathrm{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\}$	$\boldsymbol{O}$	$\boldsymbol{O}$	$\times$
$E_{\mathrm{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\}$	$\boldsymbol{O}$	$\boldsymbol{O}$	$\times$
all other sub-blocks	$\boldsymbol{O}$	$\boldsymbol{O}$	$\boldsymbol{O}$

NOTE: The i subscript ranges over $1,\ldots,m$ and the i' subscript ranges over $1,\ldots,m'$.

3.1. *The $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ Matrix and Product Restriction Implications*

The square matrix $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ has $(p+mq+m'q')^2$ entries. Therefore, a version of the Section 2 education study example involving 10,000 students is such that $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ has more than 400 million entries. However, product restrictions I, II, and III impose sparseness structures on $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$, which are summarized in Table 1. Section 4 is concerned with deriving streamlined mean field variational Bayes fitting and inference algorithms according to each of the three product restrictions. Table 1 provides a roadmap for the nature of the required results.

Under product restriction I, only the diagonal sub-blocks of $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ given by Table 1 are nonzero. These sub-blocks only have a total of $p^2+q^2m+(q')^2m'$ entries. A mean field variational Bayes algorithm that takes advantage of this sparseness will scale well to very large problems.

For product restriction II there are an additional pqm nonzero entries in $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ due to the $E_{\mathrm{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T\}$ contributions. The number of nonzero entries is still linear in m and m', but the sparsity structure is more delicate. The upcoming Result 1 is concerned with efficient approximate inference when such structure is present.

Product restriction III is particularly mild, but involves an additional $pq'm'+q'm'qm$ potentially nonzero entries in the $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})}$ matrix. If m' is moderately sized then a type of sparseness arises. Result 2 in the next section is motivated by this situation.

4. Streamlined Variational Inference

Variational inference for σ^2, $\boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}'$ is relatively straightforward and only moderately affected by the type of product restriction on the effects parameters. However, there are distinct differences among the product restrictions for updating the parameters in $\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\rm all})$ so these are treated separately in each of the next three sections. After that we treat the variance and covariance matrices component of the model.

4.1. Streamlined Variational Inference for $(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ Under Product Restriction I

Under product restriction I the variational inference updates are relatively simple and can be done using standard mean field arguments. The derivational details are given in Section S.3 of the supplementary materials.

Given current values of the q-density parameters of σ^2, $\boldsymbol{u}$, and $\boldsymbol{u}'$ the updates for the $\mathrm{q}(\boldsymbol{\beta})$ parameters are:

$$
\begin{aligned}
&\boldsymbol{b} \longleftarrow \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2} \left[\boldsymbol{y} - \underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}\left(\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)} + \boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}\right)\right\}\right] \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \end{bmatrix} \\
&\boldsymbol{B} \longleftarrow \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\boldsymbol{X} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \end{bmatrix}; \\
&\mathcal{S} \longleftarrow \textsc{SolveLeastSquares}\Big(\{\boldsymbol{b}, \boldsymbol{B}\}\Big) \\
&\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x} \text{ component of } \mathcal{S}; \\
&\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})} \longleftarrow (\boldsymbol{B}^T\boldsymbol{B})^{-1} \text{ component of } \mathcal{S}
\end{aligned} \tag{9}
$$

where the SolveLeastSquares algorithm is given by Algorithm S.2 in Section S.4 of the supplementary materials. Then, given the current values of the q-density parameters of $\boldsymbol{\beta}$, $\boldsymbol{u}'$, σ^2 and $\boldsymbol{\Sigma}$ the updates for the parameters of the $\mathrm{q}(\boldsymbol{u}_i)$, $1 \le i \le m$, have similar expressions involving the SolveLeastSquares algorithm. The updates for $\mathrm{q}(\boldsymbol{u}'_{i'})$, $1 \le i' \le m'$, are analogous.

The full set of updates is provided by Algorithm 1.

4.2. Streamlined Variational Inference for $(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ Under Product Restriction II

Under product restriction II the updates for the $\mathrm{q}(\boldsymbol{u}'_{i'})$ parameters are the same as those for product restriction I. However, streamlined updating of the $\mathrm{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters is more delicate. The problem can be embedded within the class of two-level sparse matrix problems as defined in Nolan and Wand (2020) and is encapsulated in Result 1. Note that Result 1 uses matrix sub-block notation given by (S.2) in Section S.5 of the supplementary materials. The derivation of this result is given in Section S.6 of the supplementary materials of this article.

Result 1. According to product restriction II, the mean field variational Bayes updates of $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ and each of the sub-blocks of $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ listed in the first row of Table 1, given the current values of $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}$, $1 \le i' \le m'$, are expressible as a two-level sparse matrix least squares problem of the form:

$$\left\|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}\right\|^2$$

where $\boldsymbol{b}$ and the nonzero sub-blocks of $\boldsymbol{B}$, according to the notation in (S.1) of the supplementary materials, are, for $1 \le i \le m$,

$$
\boldsymbol{b}_i \equiv \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\left\{\overset{\blacktriangle}{\boldsymbol{y}}_i - \underset{1\le i'\le m'}{\text{stack}}\left(\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}\right)\right\} \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix},
\qquad
\boldsymbol{B}_i \equiv \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\overset{\blacktriangle}{\boldsymbol{X}}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix}
$$

and

$$
\overset{\bullet}{\boldsymbol{B}}_i \equiv \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\overset{\blacktriangle}{\boldsymbol{Z}}_i \\ \boldsymbol{O} \\ \boldsymbol{M}_{\mathrm{q}(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix},
$$

with each of these matrices having $\tilde{n}_i = n_{i\bullet} + p + q$ rows. The solutions are

$$\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})} = \boldsymbol{x}_1, \quad \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$$

and

$$
\begin{aligned}
&\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)} = \boldsymbol{x}_{2,i}, \quad \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)} = \boldsymbol{A}^{22,i}, \\
&E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T\} = \boldsymbol{A}^{12,i}, \quad 1 \le i \le m,
\end{aligned}
$$

where the $\boldsymbol{x}_1$, $\boldsymbol{x}_{2,i}$, $\boldsymbol{A}^{11}$, $\boldsymbol{A}^{22,i}$, and $\boldsymbol{A}^{12,i}$ notation is given by (S.2) of the supplementary materials.

Result 1 gives rise to Algorithm 2, which provides the full set of updates of the $\mathrm{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters under product restriction II. Note that Algorithm 2 makes use of the SolveTwoLevelSparseLeastSquares algorithm from Nolan et al. (2020) and reproduced for convenience in Section S.5 of the supplementary materials.

4.3. Streamlined Variational Inference for $(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ Under Product Restriction III

Product restriction III is such that sparse least squares systems do not arise naturally in the same way as product restrictions I and II or the nested random effects models treated in Lee and Wand (2016) and Nolan et al. (2020).

Result 2 embeds the updates of the $\mathrm{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters within the class of two-level sparse matrix problems as defined in Nolan and Wand (2020) and summarized in Section S.5 of the supplementary materials. The updates are valid for any values of m and m'. If m' is moderate in size but m is possibly very large then the system is efficient in the sense that the amount of storage and computing is linear in m.

Result 2. According to product restriction III, the mean field variational Bayes updates of $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ and each of the sub-blocks of $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ in the first four rows of Table 1 is expressible as a two-level sparse matrix least squares problem of the form:

$$\left\|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}\right\|^2$$

where $\boldsymbol{b}$ and the nonzero sub-blocks of $\boldsymbol{B}$, according to the notation in (S.1) of the supplementary materials, are, for $1 \le i \le m$,

$$
\boldsymbol{b}_i \equiv \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\overset{\blacktriangle}{\boldsymbol{y}}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix},
$$

$$
\boldsymbol{B}_i \equiv \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\overset{\blacktriangle}{\boldsymbol{X}}_i & \mu_{\mathrm{q}(1/\sigma^2)}^{1/2}\overset{\blacksquare}{\boldsymbol{Z}'_i} \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} & \boldsymbol{O} \\ \boldsymbol{O} & m^{-1/2}\left(\boldsymbol{I}_{m'} \otimes \boldsymbol{M}_{\mathrm{q}((\boldsymbol{\Sigma}')^{-1})}^{1/2}\right) \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}
$$

Algorithm 1 *Mean field variational Bayes algorithm for updating the parameters of* $\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{all})$ *under product restriction I.*

Data Inputs: $(\boldsymbol{y},\boldsymbol{X})$, $\{(\overset{\blacktriangle}{\boldsymbol{y}}_i,\overset{\blacktriangle}{\boldsymbol{X}}_i,\overset{\blacktriangle}{\boldsymbol{Z}}_i):\ 1\le i\le m\}$, $\{(\overset{\blacktriangledown}{\boldsymbol{y}}_{i'},\overset{\blacktriangledown}{\boldsymbol{X}}_{i'},\overset{\blacktriangledown}{\boldsymbol{Z}'}_{i'}):\ 1\le i'\le m'\}$,
$\{(\boldsymbol{Z}_{ii'},\boldsymbol{Z}'_{ii'}):1\le i\le m,\ 1\le i'\le m'\}$

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p\times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p\times p)$ symmetric and positive definite,

q-Density Inputs: $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}$, $1\le i\le m$, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}$, $1\le i'\le m'$, $\mu_{\mathfrak{q}(1/\sigma^2)}$, $\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}(q\times q)$,
$\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}(q'\times q')$ both symmetric and positive definite.

$$\boldsymbol{b}\longleftarrow\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\left[\boldsymbol{y}-\underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}\Big(\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}+\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\Big)\right\}\right]\\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}}\end{array}\right]$$

$$\boldsymbol{B}\longleftarrow\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\boldsymbol{X}\\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\end{array}\right]\ ;\ \mathcal{S}\longleftarrow\textsc{SolveLeastSquares}\Big(\{\boldsymbol{b},\boldsymbol{B}\}\Big)$$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{x}$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow(\boldsymbol{B}^T\boldsymbol{B})^{-1}$ component of $\mathcal{S}$

For $i=1,\ldots,m$:

$$\boldsymbol{b}\longleftarrow\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\left\{\overset{\blacktriangle}{\boldsymbol{y}}_i-\overset{\blacktriangle}{\boldsymbol{X}}_i\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\underset{1\le i'\le m'}{\text{stack}}\Big(\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\Big)\right\}\\ \boldsymbol{0}\end{array}\right]$$

$$\boldsymbol{B}\longleftarrow\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\overset{\blacktriangle}{\boldsymbol{Z}}_i\\ \boldsymbol{M}^{1/2}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\end{array}\right]\ ;\ \mathcal{S}\longleftarrow\textsc{SolveLeastSquares}\Big(\{\boldsymbol{b},\boldsymbol{B}\}\Big)$$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\longleftarrow\boldsymbol{x}$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\longleftarrow(\boldsymbol{B}^T\boldsymbol{B})^{-1}$ component of $\mathcal{S}$

For $i'=1,\ldots,m'$:

$$\boldsymbol{b}\longleftarrow\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\left\{\overset{\blacktriangledown}{\boldsymbol{y}}_{i'}-\overset{\blacktriangledown}{\boldsymbol{X}}_{i'}\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\underset{1\le i\le m}{\text{stack}}\Big(\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\Big)\right\}\\ \boldsymbol{0}\end{array}\right]$$

$$\boldsymbol{B}\longleftarrow\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\overset{\blacktriangledown}{\boldsymbol{Z}'}_{i'}\\ \boldsymbol{M}^{1/2}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}\end{array}\right]\ ;\ \mathcal{S}\longleftarrow\textsc{SolveLeastSquares}\Big(\{\boldsymbol{b},\boldsymbol{B}\}\Big)$$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\longleftarrow\boldsymbol{x}$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\longleftarrow(\boldsymbol{B}^T\boldsymbol{B})^{-1}$ component of $\mathcal{S}$

Outputs: $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}$, $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}$, $\Big\{\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)},\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\Big):\ 1\le i\le m\Big\}$, $\Big\{\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})},\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\Big):1\le i'\le m'\Big\}$

$$\text{and}\quad \overset{\bullet}{\boldsymbol{B}}_i\equiv\left[\begin{array}{c}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\overset{\blacktriangle}{\boldsymbol{Z}}_i\\ \boldsymbol{O}\\ \boldsymbol{O}\\ \boldsymbol{M}^{1/2}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\end{array}\right]$$

with each of these matrices having $n_{i\bullet}+p+m'q'+q$ rows and with $\boldsymbol{B}_i$ having $p+m'q'$ columns and $\overset{\bullet}{\boldsymbol{B}}_i$ having q columns. The solutions are, with sub-matrix labeling of $\boldsymbol{x}$ and $\boldsymbol{A}^{-1}$ according to (S.2) in the supplementary materials,

$$\begin{aligned}
\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}&=\ \text{first } p \text{ rows of } \boldsymbol{x}_1,\\
\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}&=\ \text{top left } p\times p \text{ sub-block of } \boldsymbol{A}^{11},\\
\underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})&=\text{subsequent } (m'q')\times 1 \text{ entries of}\\
&\qquad \boldsymbol{x}_1\ \text{ following } \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})},\\
E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}&=\ \text{subsequent } p\times q'\\
&\qquad \text{sub-blocks of } \boldsymbol{A}^{11} \text{ to the right of } \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})},\\
\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}&=\ \text{subsequent } q'\times q' \text{ diagonal sub-blocks of } \boldsymbol{A}^{11}\\
&\qquad \text{following } \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})},\ \ 1\le i'\le m',\\
\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}&=\boldsymbol{x}_{2,i},\ \ \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}=\boldsymbol{A}^{22,i},\\
E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}&=\text{first } p \text{ rows of } \boldsymbol{A}^{12,i}\\
\text{and}\quad \underset{1\le i'\le m'}{\text{stack}}\Big([E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}]^T\Big)&\\
&=\ \text{remaining } m'q' \text{ rows of } \boldsymbol{A}^{12,i},\ \ 1\le i\le m,
\end{aligned}$$

where the $\boldsymbol{x}_1$, $\boldsymbol{x}_{2,i}$, $\boldsymbol{A}^{11}$, $\boldsymbol{A}^{22,i}$, and $\boldsymbol{A}^{12,i}$ notation is given by (S.2) of the supplementary materials.

Figure 1 provides visualization of the strategy used by Result 2. For simplicity, the values of p, q, q' and $n_{ii'}$ are all set to 1 and m' is set to 2. Each panel shows an image plot representation of the matrix $\boldsymbol{B}$ according to the sparse two-level form given by (S.1) of the supplementary materials but with the $\boldsymbol{B}_i$ and $\overset{\bullet}{\boldsymbol{B}}_i$ sub-blocks specific to Result 2. The white regions correspond to the two-level sparsity due to

Algorithm 2 *Mean field variational Bayes algorithm for updating the parameters of* $\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ *under product restriction II.*

Data Inputs: $\{(\overset{\blacktriangle}{\boldsymbol{y}}_i, \overset{\blacktriangle}{\boldsymbol{X}}_i, \overset{\blacktriangle}{\boldsymbol{Z}}_i) : 1 \le i \le m\}$, $\{(\overset{\blacktriangledown}{\boldsymbol{y}}_{i'}, \overset{\blacktriangledown}{\boldsymbol{X}}_{i'}, \overset{\blacktriangledown}{\boldsymbol{Z}'}_{i'}) : 1 \le i' \le m'\}$,
$\{(\boldsymbol{Z}_{ii'}, \boldsymbol{Z}'_{ii'}) : 1 \le i \le m,\ 1 \le i' \le m'\}$

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p \times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p \times p)$ symmetric and positive definite.

q-Density Inputs: $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}$, $1 \le i' \le m'$, $\mu_{\mathsf{q}(1/\sigma^2)}$, $\boldsymbol{M}_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})}(q \times q)$,
$\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}')^{-1})}(q' \times q')$ both symmetric and positive definite.

For $i = 1, \ldots, m$:

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathsf{q}(1/\sigma^2)} \left\{ \overset{\blacktriangle}{\boldsymbol{y}}_i - \underset{1 \le i' \le m'}{\text{stack}} \left(\boldsymbol{Z}'_{ii'} \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})} \right) \right\} \\ m^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix} ; \quad \boldsymbol{B}_i \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathsf{q}(1/\sigma^2)} \overset{\blacktriangle}{\boldsymbol{X}}_i \\ m^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix}$$

$$\overset{\bullet}{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathsf{q}(1/\sigma^2)} \overset{\blacktriangle}{\boldsymbol{Z}}_i \\ \boldsymbol{O} \\ \boldsymbol{M}^{1/2}_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}$$

$$\mathcal{S} \longleftarrow \textsc{SolveTwoLevelSparseLeastSquares}\Big(\big\{ (\boldsymbol{b}_i, \boldsymbol{B}_i, \overset{\bullet}{\boldsymbol{B}}_i) : 1 \le i \le m \big\} \Big)$$

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}$

For $i = 1, \ldots, m$:

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}$

$E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}$

For $i' = 1, \ldots, m'$:

$$\boldsymbol{b} \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathsf{q}(1/\sigma^2)} \left\{ \overset{\blacktriangledown}{\boldsymbol{y}}_{i'} - \overset{\blacktriangledown}{\boldsymbol{X}}_{i'} \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \underset{1 \le i \le m}{\text{stack}} \left(\boldsymbol{Z}_{ii'} \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} \right) \right\} \\ \boldsymbol{0} \end{bmatrix}$$

$$\boldsymbol{B} \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathsf{q}(1/\sigma^2)} \overset{\blacktriangledown}{\boldsymbol{Z}'}_{i'} \\ \boldsymbol{M}^{1/2}_{\mathsf{q}((\boldsymbol{\Sigma}')^{-1})} \end{bmatrix} ; \quad \mathcal{S} \longleftarrow \textsc{SolveLeastSquares}\Big(\{\boldsymbol{b}, \boldsymbol{B}\} \Big)$$

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})} \longleftarrow \boldsymbol{x}$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}'_{i'})} \longleftarrow (\boldsymbol{B}^T\boldsymbol{B})^{-1}$ component of $\mathcal{S}$

Outputs: $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}$, $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}$, $\left\{ \left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} \right) : 1 \le i \le m \right\}$, $\left\{ \left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}'_{i'})} \right) : 1 \le i' \le m' \right\}$,
$\left\{ E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T : 1 \le i \le m \right\}$

the block diagonal positioning of the $\overset{\bullet}{\boldsymbol{B}}_i$, $1 \le i \le m$. The grey regions also indicate entries, and have additional block diagonal formations, but which do not contribute to the two-level sparsity. For moderate m' and large m the black/grey block on the left is small relative to the remainder of the matrix. The SolveTwoLevelSparseLeastSquares algorithm, listed as Algorithm S.3 in Section S.5 of the supplementary materials, affords efficient calculation of the variational inference updates for m potentially very large.

An interesting future research problem concerns taking advantage of the sparseness apparent in the grey regions of the $\boldsymbol{B}$ matrices displayed in Figure 1. This is a much more subtle pattern of sparseness compared with the two-level sparse structure corresponding to the white regions in Figure 1 and accounting for it would require significant additional algebraic analysis.

Algorithm 3 is a proceduralization of Result 2 and delivers the full set of updates of the $\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters under product restriction III.

4.4. Variational Inference for σ^2, Σ and Σ'

Given the current values of the $\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters, the updates of the parameters of $\mathsf{q}(\sigma^2)$, $\mathsf{q}(\boldsymbol{\Sigma})$ and $\mathsf{q}(\boldsymbol{\Sigma}')$ are relatively simple. For example, σ^2 has the Inverse χ^2 prior as given by (3) then standard mean field variational Bayes arguments (e.g., Bishop 2006, sec. 10.1–10.3) lead to $\xi_{\mathsf{q}(\sigma^2)} = \xi_{\sigma^2} + n_{\bullet\bullet}$ and

$$\begin{aligned} \lambda_{\mathsf{q}(\sigma^2)} &= \lambda_{\sigma^2} + E_{\mathsf{q}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u}_{\text{all}}\|^2 \\ &= \lambda_{\sigma^2} + \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{\text{all}})}\|^2 \\ &\quad + \text{tr}\big([\boldsymbol{X}\ \boldsymbol{Z}] \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})} \big). \end{aligned}$$

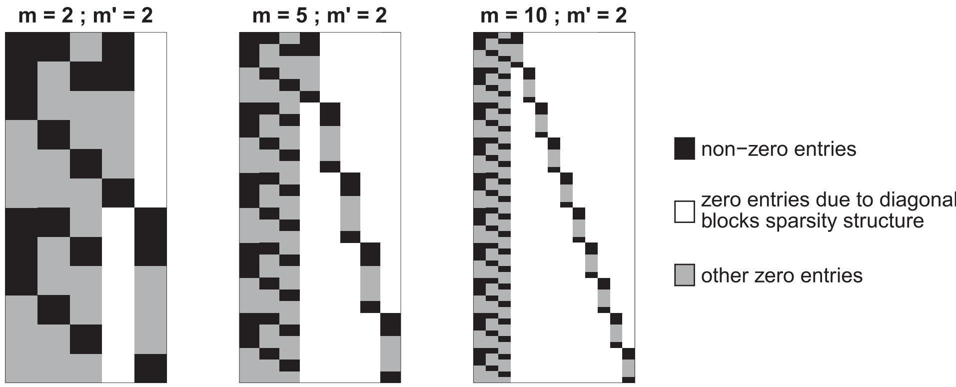

Figure 1. Image plot representation of the two-level sparse matrix $\boldsymbol{B}$ with generic form given by (S.1) of the supplementary materials and with sub-blocks as defined in Result 2. The dimension variables are $p = q = q' = n_{ii'} = 1$, $m \in \{2, 5, 10\}$ and $m' = 2$. Black indicates nonzero entries of $\boldsymbol{B}$. White indicates zero entries corresponding to the diagonal blocks sparsity structure of $\boldsymbol{B}$. The gray regions also correspond to zero entries but which do not contribute to two-level sparse structure.

Under product restriction I the trace term reduces to

$$\mathrm{tr}\left\{[\boldsymbol{X}\ \boldsymbol{Z}]^T[\boldsymbol{X}\ \boldsymbol{Z}]\,\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}\right\} = \sum_{i=1}^{m}\sum_{i'=1}^{m'}\Big\{\mathrm{tr}(\boldsymbol{X}_{ii'}^T\boldsymbol{X}_{ii'}\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})}) + \mathrm{tr}(\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}_{ii'}\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)}) + \mathrm{tr}(\boldsymbol{Z}_{ii'}'^T\boldsymbol{Z}_{ii'}'\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_{i'}')})\Big\}.$$

For product restrictions II and III additional terms are present due to nonzero cross-expectations and is reflected in the $\lambda_{\mathrm{q}(\sigma^2)}$ updates in Algorithm 4 given in the next section.

The updates for the parameters of $\mathrm{q}(\boldsymbol{\Sigma})$ and $\mathrm{q}(\boldsymbol{\Sigma}')$ uses analogous arguments, and this is also reflected in the $\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})}$ and $\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')}$ updates of Algorithm 4.

4.5. Full Streamlined Mean Field Variational Algorithm

We are now ready to list a full streamlined mean field variational inference algorithm, listed as Algorithm 4, that accounts for any of product restrictions I, II, or III. It also allows for the covariance matrix prior specification to be (3) or (4).

Throughout this article we confine discussion to the Gaussian response version of the linear mixed model with crossed random effects. Item response theory and Rasch analysis models, which enjoy widespread use in psychometrics, have random effects structures similar to those given by (1). They usually involve different conditional response distributions such as those corresponding to multivariate binary and multivariate categorical data. However, the streamlined variational inference challenges arising in random effect structures are independent of the likelihood. The variational message passing approach to variational inference (e.g., Wand 2017; Nolan et al. 2020) formalizes this separation via notions such as factor graph fragments. The upshot is that Results 1 and 2 are still relevant to non-Gaussian crossed random effects models such as the psychometrics versions just mentioned.

5. Performance Assessment and Comparison

Any set of statistical methods for a particular problem can be assessed and compared on various criteria such as ease of implementation, time to compute and various measures of statistical accuracy. In this section we focus on accuracy in terms of how close variational approximate posterior density functions are to their exact counterparts and computational speed. The second of these assessments and comparisons allows appreciation for the scalability of competing approaches to very large mixed models with crossed random effects.

5.1. Accuracy Assessment and Comparison

We ran a simulation study to compare and assess the accuracy performance of the three mean field variational inference schemes. The study involved simulating 100 replications of data from a version of the crossed random effects model (1). The dimension variables were set to be:

$$m = 100,\quad m' = 20,\quad n_{ii'} = 10,\quad \text{and}\quad p = q = q' = 2.$$

The true values of the parameters from which the data were generated are

$$\boldsymbol{\beta}_{\mathrm{true}} = \begin{bmatrix} 0.58 \\ 1.89 \end{bmatrix},\quad \sigma^2_{\mathrm{true}} = 0.3,\quad \boldsymbol{\Sigma}_{\mathrm{true}} = \begin{bmatrix} 0.46 & -0.19 \\ -0.19 & 0.17 \end{bmatrix},\ \text{and}$$
$$\boldsymbol{\Sigma}'_{\mathrm{true}} = \begin{bmatrix} 0.3 & -0.12 \\ -0.12 & 0.25 \end{bmatrix}. \qquad (10)$$

Each of the $\boldsymbol{X}_{ii'}, \boldsymbol{Z}_{ii'}, \boldsymbol{Z}'_{ii'}$, $1 \le i \le 100$, $1 \le i' \le 20$, were 10×2 matrices with a column of ones and a column of predictor values generated to be independent and uniformly on the unit interval.

The priors on σ^2, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ were of (4). The hyperparameter values were $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{0}$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^{10}\boldsymbol{I}$, $\nu_{\sigma^2} = 1$, $\nu_{\boldsymbol{\Sigma}} = \nu_{\boldsymbol{\Sigma}'} = 2$ and $s_{\sigma^2} = s_{\boldsymbol{\Sigma},1} = s_{\boldsymbol{\Sigma},2} = s_{\boldsymbol{\Sigma}',1} = s_{\boldsymbol{\Sigma}',2} = 10^5$.

For each replication we obtained approximate posterior density functions for all model parameters and random effects using both mean field variational Bayes and Markov chain Monte Carlo. The mean field variational Bayes approximations were obtained by running Algorithm 4 with each of product restrictions I, II, and III. The number of iterations was fixed at 500. Markov chain Monte Carlo approximate density functions were obtained using the package `rstan` (Stan Development Team 2021) within the R language (R Core Team 2019). One thousand warm-up samples were generated, followed by another 1000 samples retained for approximate inference. Kernel density

Algorithm 3 *Mean field variational Bayes algorithm for updating the parameters of* $\mathrm{q}(\boldsymbol{\beta}, \boldsymbol{u}_{all})$ *under product restriction III.*

Data Inputs: $\left\{ \left(\overset{\blacktriangle}{\boldsymbol{y}}_i, \overset{\blacktriangle}{\boldsymbol{X}}_i, \overset{\blacktriangle}{\boldsymbol{Z}}_i, \overset{\blacksquare}{\boldsymbol{Z}'}_i \right) : 1 \le i \le m \right\}$

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p \times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p \times p)$ symmetric and positive definite,

q-Density Inputs: $\mu_{\mathrm{q}(1/\sigma^2)} > 0$, $\boldsymbol{M}_{\mathrm{q}(\boldsymbol{\Sigma}^{-1})}(q \times q)$, $\boldsymbol{M}_{\mathrm{q}((\boldsymbol{\Sigma}')^{-1})}(q' \times q')$ symmetric and positive definite.

For $i = 1, \ldots, m$:

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2} \overset{\blacktriangle}{\boldsymbol{y}}_i \\ m^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}; \quad \boldsymbol{B}_i \longleftarrow \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2} \overset{\blacktriangle}{\boldsymbol{X}}_i & \mu_{\mathrm{q}(1/\sigma^2)}^{1/2} \overset{\blacksquare}{\boldsymbol{Z}'_i} \\ m^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} & \boldsymbol{O} \\ \boldsymbol{O} & m^{-1/2} \left(\boldsymbol{I}_{m'} \otimes \boldsymbol{M}_{\mathrm{q}((\boldsymbol{\Sigma}')^{-1})}^{1/2} \right) \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}$$

$$\overset{\bullet}{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \mu_{\mathrm{q}(1/\sigma^2)}^{1/2} \overset{\blacktriangle}{\boldsymbol{Z}}_i \\ \boldsymbol{O} \\ \boldsymbol{O} \\ \boldsymbol{M}_{\mathrm{q}(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix}.$$

$\mathcal{S} \longleftarrow$ SolveTwoLevelSparseLeastSquares$\left(\left\{ (\boldsymbol{b}_i, \boldsymbol{B}_i, \overset{\bullet}{\boldsymbol{B}}_i) : 1 \le i \le m \right\} \right)$

$\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})} \longleftarrow$ first p rows of $\boldsymbol{x}_1$ component of $\mathcal{S}$

$\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})} \longleftarrow$ top left $p \times p$ sub-block of $\boldsymbol{A}^{11}$ component of $\mathcal{S}$

$i_{\mathrm{stt}} \longleftarrow p + 1$

For $i' = 1, \ldots, m'$:

> $i_{\mathrm{end}} \longleftarrow i_{\mathrm{stt}} + q' - 1$
>
> $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})} \longleftarrow$ sub-vector of $\boldsymbol{x}_1$ component of $\mathcal{S}$ with entries i_{stt} to i_{end}
>
> $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}'_{i'})} \longleftarrow$ diagonal sub-block of $\boldsymbol{A}^{11}$ component of $\mathcal{S}$ with rows i_{stt} to i_{end} and columns i_{stt} to i_{end}
>
> $E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\} \longleftarrow$ sub-block of $\boldsymbol{A}^{11}$ component of $\mathcal{S}$ with rows 1 to p and columns i_{stt} to i_{end}
>
> $i_{\mathrm{stt}} \longleftarrow i_{\mathrm{end}} + 1$

For $i = 1, \ldots, m$:

> $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}$; $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}$
>
> $E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T\} \longleftarrow$ sub-matrix of $\boldsymbol{A}^{12,i}$ component of $\mathcal{S}$ with rows 1 to p
>
> $\boldsymbol{\Omega} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}$; $i_{\mathrm{stt}} \longleftarrow p + 1$
>
> For $i' = 1, \ldots, m'$:
>
> > $i_{\mathrm{end}} \longleftarrow i_{\mathrm{stt}} + q' - 1$
> >
> > $E_q\{(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\} \longleftarrow$ sub-matrix of $\boldsymbol{\Omega}^T$ with columns i_{stt} to i_{end}
> >
> > $i_{\mathrm{stt}} \longleftarrow i_{\mathrm{end}} + 1$

Outputs: $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})}, \left\{ \left(\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)} \right) : 1 \le i \le m \right\}, \left\{ \left(\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}, \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}'_{i'})} \right) : 1 \le i' \le m' \right\}$,
$\left\{ E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T : 1 \le i \le m \right\}, \left\{ E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T : 1 \le i' \le m' \right\}$,
$\left\{ E_{\mathrm{q}}\{(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\} : 1 \le i \le m,\ 1 \le i' \le m' \right\}$

estimation, with direct plug-in bandwidth selection (e.g., Wand and Jones 1995, sec. 3.6.1), was used to obtain approximate posterior density functions.

Figure 2 compares the approximations for the posterior distributions of the two entries of $\boldsymbol{\beta}$. We denote these entries as β_0, the fixed effects intercept, and β_1, the fixed effects slope.

Algorithm 4 *Mean field variational Bayes algorithm for determining the optimal* q*-density parameters in the Bayesian crossed random effects model under either product restriction I, II or III.*

Data Inputs: $\boldsymbol{y}_{ii'},\ \boldsymbol{X}_{ii'},\ \boldsymbol{Z}_{ii'},\ \boldsymbol{Z}'_{ii'},\quad 1 \le i \le m,\ 1 \le i' \le m'$.

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p \times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p \times p)$ symmetric and positive definite.

If priors (3): $\xi_{\sigma^2}, \lambda_{\sigma^2} > 0, \xi_{\boldsymbol{\Sigma}} > 2(q-1), \xi_{\boldsymbol{\Sigma}'} > 2(q'-1)$,

$\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}, \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}'}$ positive definite.

If priors (4): $\nu_{\sigma^2}, s_{\sigma^2}, \nu_{\boldsymbol{\Sigma}}, \nu_{\boldsymbol{\Sigma}'}, s_{\boldsymbol{\Sigma},1}, \ldots, s_{\boldsymbol{\Sigma},q}, s_{\boldsymbol{\Sigma}',1}, \ldots, s_{\boldsymbol{\Sigma}',q'} > 0$.

Product Restriction Input: Specification of product restriction I, II or III.

$$\overset{\blacktriangle}{\boldsymbol{y}}_i \longleftarrow \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{y}_{ii'}),\ \overset{\blacktriangle}{\boldsymbol{X}}_i \longleftarrow \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{X}_{ii'}),\ \overset{\blacktriangle}{\boldsymbol{Z}}_i \longleftarrow \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{Z}_{ii'}),\quad 1 \le i \le m.$$

If product restriction III then: $\overset{\blacksquare}{\boldsymbol{Z}'_i} \longleftarrow \underset{1\le i'\le m'}{\text{blockdiag}}(\boldsymbol{Z}'_{ii'}),\quad 1 \le i \le m$

If product restriction I or II then: $\overset{\blacktriangledown}{\boldsymbol{y}}_{i'} \longleftarrow \underset{1\le i\le m}{\text{stack}}(\boldsymbol{y}_{ii'}),\ \overset{\blacktriangledown}{\boldsymbol{X}}_{i'} \longleftarrow \underset{1\le i\le m}{\text{stack}}(\boldsymbol{X}_{ii'})$,

$$\overset{\blacktriangledown}{\boldsymbol{Z}'}_{i'} \longleftarrow \underset{1\le i\le m}{\text{stack}}(\boldsymbol{Z}'_{ii'}),\quad 1 \le i' \le m'.$$

If product restriction I then: $\boldsymbol{y} \longleftarrow \underset{1\le i\le m}{\text{stack}}(\overset{\blacktriangle}{\boldsymbol{y}}_i),\ \boldsymbol{X} \longleftarrow \underset{1\le i\le m}{\text{stack}}(\overset{\blacktriangle}{\boldsymbol{X}}_i)$.

If priors (3)

$$\xi_{\mathrm{q}(\sigma^2)} \longleftarrow \xi_{\sigma^2} + n_{\bullet\bullet}\quad ;\quad \xi_{\mathrm{q}(\boldsymbol{\Sigma})} \longleftarrow \xi_{\boldsymbol{\Sigma}} + m,\quad ;\quad \xi_{\mathrm{q}(\boldsymbol{\Sigma}')} \longleftarrow \xi_{\boldsymbol{\Sigma}'} + m'$$

If priors (4)

initialize: $\mu_{\mathrm{q}(1/a_{\sigma^2})} > 0, \boldsymbol{M}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}, \boldsymbol{M}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1})}$ positive definite.

$$\xi_{\mathrm{q}(\sigma^2)} \longleftarrow \nu_{\sigma^2} + n_{\bullet\bullet}\quad ;\quad \xi_{\mathrm{q}(\boldsymbol{\Sigma})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + 2q - 2 + m,\quad ;\quad \xi_{\mathrm{q}(\boldsymbol{\Sigma}')} \longleftarrow \nu_{\boldsymbol{\Sigma}'} + 2q' - 2 + m'$$

$$\xi_{\mathrm{q}(a_{\sigma^2})} \longleftarrow \nu_{\sigma^2} + 1\quad ;\quad \xi_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + q\quad ;\quad \xi_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})} \longleftarrow \nu_{\boldsymbol{\Sigma}'} + q'$$

Initialize: $\mu_{\mathrm{q}(1/\sigma^2)} > 0, \boldsymbol{M}_{\mathrm{q}(\boldsymbol{\Sigma}^{-1})}, \boldsymbol{M}_{\mathrm{q}((\boldsymbol{\Sigma}')^{-1})}$ positive definite.

Cycle:

If prod. restrict. I: call Algorithm 1 to update $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}$ and relevant $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}$ blocks

If prod. restrict. II: call Algorithm 2 to update $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}$ and relevant $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}$ blocks

If prod. restrict. III: call Algorithm 3 to update $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}$ and relevant $\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta},\boldsymbol{u}_{\mathrm{all}})}$ blocks

If priors (3): $\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \lambda_{\sigma^2}\quad ;\quad \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}\quad ;\quad \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')} \longleftarrow \boldsymbol{\Lambda}_{\boldsymbol{\Sigma}'}$

If priors (4): $\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \mu_{\mathrm{q}(1/a_{\sigma^2})}\quad ;\quad \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{M}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\quad ;\quad \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')} \longleftarrow \boldsymbol{M}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1})}$

For $i = 1, \ldots, m$:

For $i' = 1, \ldots, m'$:

$$\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \lambda_{\mathrm{q}(\sigma^2)} + \left\| \boldsymbol{y}_{ii'} - \boldsymbol{X}_{ii'}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)} - \boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})} \right\|^2$$

$$\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \lambda_{\mathrm{q}(\sigma^2)} + \mathrm{tr}(\boldsymbol{X}_{ii'}^T\boldsymbol{X}_{ii'}\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})}) + \mathrm{tr}(\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}_{ii'}\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)}) + \mathrm{tr}(\boldsymbol{Z}_{ii'}'^T\boldsymbol{Z}'_{ii'}\boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}'_{i'})})$$

If product restriction II or III:

$$\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \lambda_{\mathrm{q}(\sigma^2)} + 2\,\mathrm{tr}\big[\boldsymbol{Z}_{ii'}^T\boldsymbol{X}_{ii'}E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T\}\big]$$

If product restriction III:

$$\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \lambda_{\mathrm{q}(\sigma^2)} + 2\,\mathrm{tr}\big[\boldsymbol{Z}_{ii'}'^T\boldsymbol{X}_{ii'}E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\}\big]$$

$$\lambda_{\mathrm{q}(\sigma^2)} \longleftarrow \lambda_{\mathrm{q}(\sigma^2)} + 2\,\mathrm{tr}\big[\boldsymbol{Z}_{ii'}'^T\boldsymbol{Z}_{ii'}E_{\mathrm{q}}\{(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\}\big]$$

continued on a subsequent page ...

Algorithm 4 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

For $i = 1, \ldots, m$:

$$\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})} + \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)}$$

For $i' = 1, \ldots, m'$:

$$\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')} \longleftarrow \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')} + \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}^T + \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}'_{i'})}$$

$$\mu_{\mathrm{q}(1/\sigma^2)} \leftarrow \xi_{\mathrm{q}(\sigma^2)}/\lambda_{\mathrm{q}(\sigma^2)} \quad ; \quad \boldsymbol{M}_{\mathrm{q}(\boldsymbol{\Sigma}^{-1})} \longleftarrow (\xi_{\mathrm{q}(\boldsymbol{\Sigma})} - q + 1)\,\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})}^{-1}$$

$$\boldsymbol{M}_{q((\boldsymbol{\Sigma}')^{-1})} \longleftarrow (\xi_{\mathrm{q}(\boldsymbol{\Sigma}')} - q' + 1)\,\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')}^{-1}$$

If priors (4):

$$\lambda_{\mathrm{q}(a_{\sigma^2})} \longleftarrow \mu_{\mathrm{q}(1/\sigma^2)} + 1/(\nu_{\sigma^2} s_{\sigma^2}^2) \quad ; \quad \mu_{\mathrm{q}(1/a_{\sigma^2})} \longleftarrow \xi_{\mathrm{q}(a_{\sigma^2})}/\lambda_{\mathrm{q}(a_{\sigma^2})}$$

$$\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} \longleftarrow \mathrm{diag}\big\{\mathrm{diagonal}\big(\boldsymbol{M}_{\mathrm{q}(\boldsymbol{\Sigma}^{-1})}\big)\big\} + \{\nu_{\boldsymbol{\Sigma}}\mathrm{diag}(s_{\boldsymbol{\Sigma},1}^2, \ldots, s_{\boldsymbol{\Sigma},q}^2)\}^{-1}$$

$$\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})} \longleftarrow \mathrm{diag}\big\{\mathrm{diagonal}\big(\boldsymbol{M}_{\mathrm{q}((\boldsymbol{\Sigma}')^{-1})}\big)\big\} + \{\nu_{\boldsymbol{\Sigma}'}\mathrm{diag}(s_{\boldsymbol{\Sigma}',1}^2, \ldots, s_{\boldsymbol{\Sigma}',q'}^2)\}^{-1}$$

$$\boldsymbol{M}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})} \longleftarrow \xi_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}^{-1} \quad ; \quad \boldsymbol{M}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1})} \longleftarrow \xi_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})}\boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})}^{-1}$$

Outputs: $\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{\beta})}, \big\{\big(\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}_i)}\big) : 1 \le i \le m\big\}, \Big\{\Big(\boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})}, \boldsymbol{\Sigma}_{\mathrm{q}(\boldsymbol{u}'_{i'})}\Big) : 1 \le i' \le m'\Big\},$

$\xi_{\mathrm{q}(\sigma^2)}, \lambda_{\mathrm{q}(\sigma^2)}, \xi_{\mathrm{q}(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma})}, \xi_{\mathrm{q}(\boldsymbol{\Sigma}')}, \boldsymbol{\Lambda}_{\mathrm{q}(\boldsymbol{\Sigma}')}$.

If product restriction II or III add: $\big\{E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})^T : 1 \le i \le m\big\}$.

If product restriction III add: $\Big\{E_{\mathrm{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T : 1 \le i' \le m'\Big\}$.

and $\Big\{E_{\mathrm{q}}\{(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}_i)})(\boldsymbol{u}_{i'} - \boldsymbol{\mu}_{\mathrm{q}(\boldsymbol{u}'_{i'})})^T\} : 1 \le i \le m,\ 1 \le i' \le m'\Big\}$.

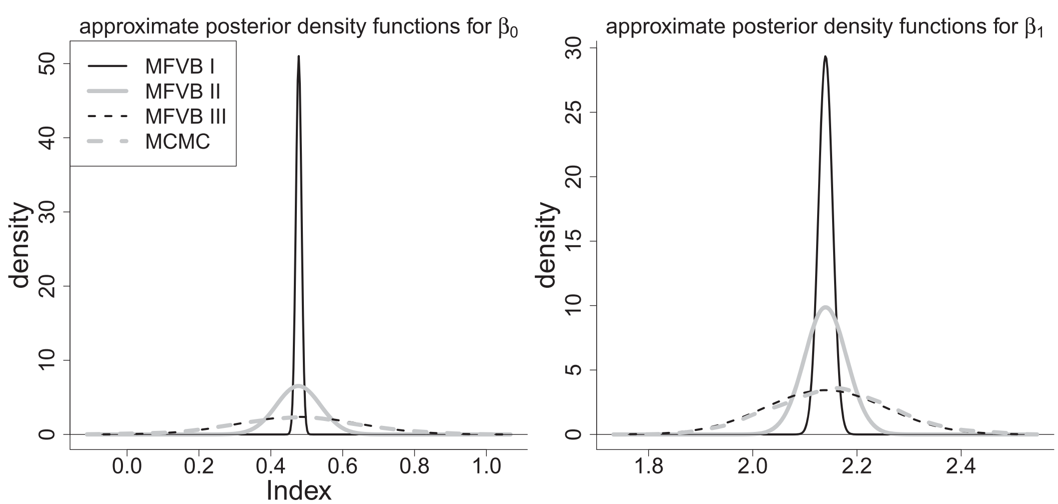

Figure 2. Approximate posterior density functions for β_0 and β_1, according to three different mean field variational Bayes (MFVB) schemes and Markov chain Monte Carlo (MCMC), for the first replication of the simulation study. The legend uses the abbreviation "MFVB I" for the mean field variational Bayes according to product restriction I. Similar abbreviations are used for the other product restrictions.

The difference between the three variational approximations is quite striking. For product restriction I the posterior variances are much too low, due to posterior correlations between the entries of $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\boldsymbol{u}'$ being set to zero. However, the product restriction III leads to very good concordance with the Markov chain Monte Carlo posterior densities. The density functions for product restriction II have intermediate approximation quality, but appear to be closer to those of product restriction III than those of product restriction I.

In Figure 3 we provide a summary of the relative performance of product restrictions I, II, and III for all model parameters and entries of the first three $\boldsymbol{u}_i$ and $\boldsymbol{u}'_{i'}$ vectors using side-by-side boxplots of estimates of the following accuracy score for a generic target θ:

$$\text{accuracy} \equiv 100\left\{1 - \tfrac{1}{2}\int_{-\infty}^{\infty}\big|\mathrm{q}(\theta) - \mathfrak{p}(\theta|\boldsymbol{y})\big|\,d\theta\right\}\%. \quad (11)$$

Note that $0\% \le \text{accuracy} \le 100\%$ with a score of 100% if $\mathrm{q}(\theta)$ and $\mathfrak{p}(\theta|\boldsymbol{y})$ perfectly coincide and a score of 0% if there have no overlapping mass. In practice $\mathfrak{p}(\theta|\boldsymbol{y})$ is replaced by a kernel density estimate based on a large Markov chain Monte Carlo sample. Depending on tractability, either $\mathrm{q}(\theta)$ is available

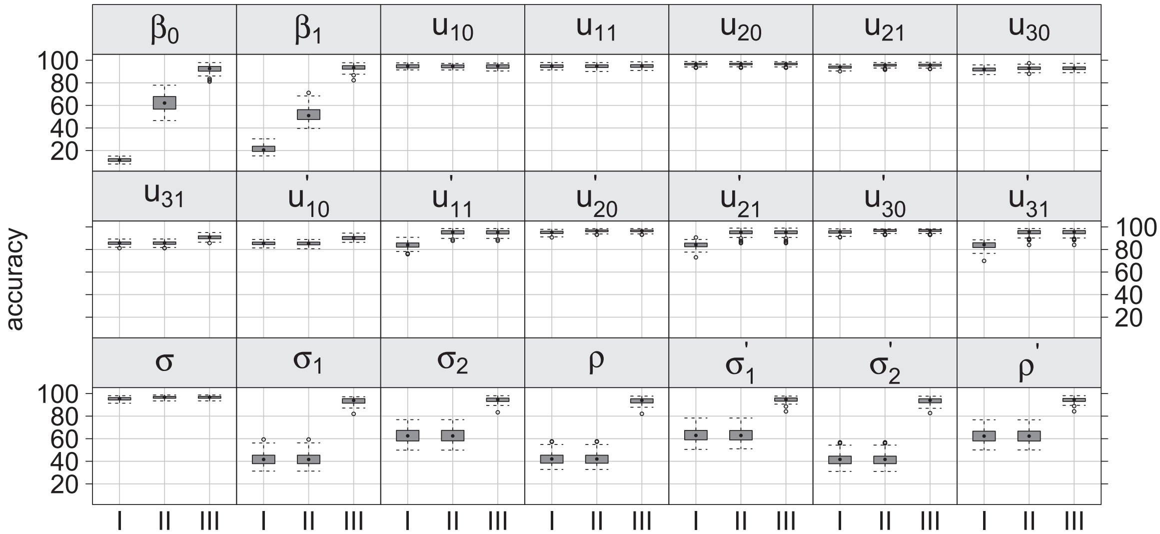

Figure 3. Side-by-side boxplots for the accuracy scores for 21 parameters and random effects from the simulation study, with accuracy defined according to (11). Each panel corresponds to a separate parameter or random effect and contains side-by-side boxplots for product restrictions I, II, and III.

in closed form or it can be estimated from a large Monte Carlo sample from the distribution corresponding to $\mathfrak{q}(\theta)$.

Apart from the fixed effects parameters β_0 and β_1 the parameters monitored in Figure 3 are the error standard deviation σ, the standard deviation and correlation parameters corresponding to the random effects covariance matrix $\boldsymbol{\Sigma}$:

$$\sigma_1 \equiv \sqrt{(\boldsymbol{\Sigma})_{11}}, \quad \sigma_2 \equiv \sqrt{(\boldsymbol{\Sigma})_{22}} \quad \text{and} \quad \rho \equiv (\boldsymbol{\Sigma})_{12}/(\sigma_1\sigma_2)$$

and similar parameters for the random effects covariance matrix $\boldsymbol{\Sigma}'$. The random effects in Figure 3 have notation as given by

$$\boldsymbol{u}_i = \begin{bmatrix} u_{i0} \\ u_{i1} \end{bmatrix}, \ 1 \le i \le 3, \quad \text{and} \quad \boldsymbol{u}'_{i'} = \begin{bmatrix} u_{i'0} \\ u_{i'1} \end{bmatrix}, \ 1 \le i' \le 3.$$

From Figure 3 we see that the biggest discrepancies across the three product restrictions are for the fixed effects parameters β_0 and β_1, which is in keeping with Figure 2. Inferential accuracy for the covariance matrix parameters is very good for all product restrictions and is excellent for product restriction III. For the $\boldsymbol{u}_i$ entries the accuracy of product restriction I is lower due to its ignorance of the posterior correlations between distinct $\boldsymbol{u}_i$ vectors. Product restrictions II and III allow for such correlation and excellent accuracy ensues. However, for the $\boldsymbol{u}'_{i'}$ vectors product restriction II sacrifices handling of the corresponding posterior correlations and the drop in accuracy is quite pronounced.

Since product restriction III is the clear winner in terms of accuracy, we show the mean field variational Bayes approximate density estimates for the product restriction in comparison with Markov chain Monte Carlo for the first replication in Figure 4. The parameters and random effects subsets are the same as those used in Figure 3. Accuracy scores are also shown and, for this dataset, is always 92% or higher. The boxplots in Figure 3 indicate that excellent accuracy is typical for this particular simulation setting.

The excellent accuracy under product restriction III is tied to the orthogonality between $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')$ and $(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}')$ from likelihood theory, h-likelihood theory and best prediction for the frequentist version of (1). Section 3.1 of Menictas and Wand (2013) provides a detailed account of this phenomenon for a similar model. The approximately non-informative priors used in this section's empirical studies imply that the approximate Bayesian inference is close to what would be obtained using frequentist paradigms. Since the product density forms of product restriction III separate orthogonal quantities, there is little loss in accuracy compared with the unrestricted case. On the other hand, there is no such orthogonality within the components of $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')$. Hence, product restrictions I and II pay a price for imposing their product density constraints.

5.2. Speed Assessment and Comparison

We ran another simulation study that recorded computing times for data generated according to the model as in the previous section's simulation study—but with increasing crossed random effects dimensions. Specifically, the data were generated according to (10) with $n_{ii'} = 10$ but with

$$m \in \{100, 200, 400, 800\} \quad \text{and} \quad m' = m/5.$$

We then simulated 10 replications of the data for each (m, m') combination and recorded the computational times for fitting via mean field variational Bayes with product restrictions II and III.

The mean field variational Bayes computations were performed using Algorithm 4, with calls to Algorithms 2 and 3 as well as Algorithms S.2 and S.3 of the supplementary materials. All five algorithms were implemented in the fast Fortran 77 language. The number of mean field variational Bayes iterations was fixed at 100. All computations were carried out on the third author's MacBook Air laptop, which has a 2.2 gigahertz processor and 8 gigabytes of random access memory.

Table 2 lists the average and standard deviation times in seconds.

Table 2 shows that mean field variational Bayes with product restriction II scales very well to large crossed random effects problems with less than a minute required for the largest $(m, m') = (800, 160)$ case and less than 10 sec required for

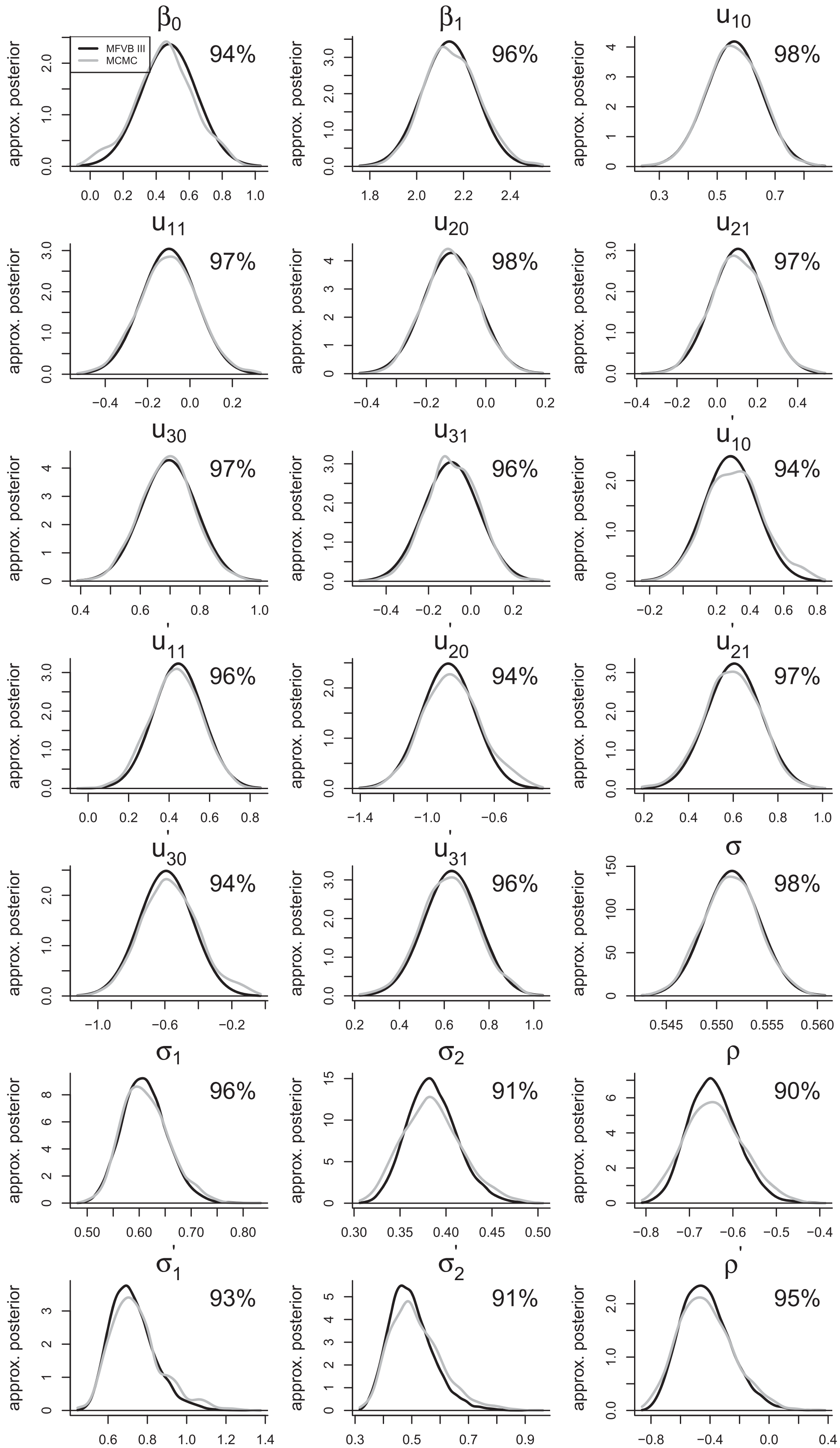

Figure 4. Approximate posterior density functions for the 21 parameters and random effects for the first replication of the simulation study. The blue curves are posterior density functions obtained using mean field variational Bayes with product restriction III and the orange curves are based on Markov chain Monte Carlo. The accuracy percentages are defined according to (11).

Table 2. Average (standard deviation) time in seconds for each method, in the speed assessment study.

(m, m')	MFVB II	MFVB III
(100,20)	0.267 (0.0267)	4.93 (0.082)
(200,40)	1.44 (0.0996)	66.8 (1.40)
(400,80)	8.92 (0.587)	1130 (8.48)
(800,160)	54.7 (1.54)	21300 (41.0)

NOTE: "MFVB II" is short for the mean field variational Bayes according to product restriction II and "MFVB III" is defined similarly.

the second largest $(m, m') = (400, 80)$ situation. The highly accurate Mean field variational Bayes with product restriction III computes in a few seconds for $(m, m') = (100, 20)$ and about a minute for $(m, m') = (200, 40)$. But eventually it gets affected by the quadratic dependence on (m, m') and the average computing time up to about 6 hr for $(m, m') = (800, 160)$, which is about 400 times slower than for product restriction II. As we have seen in Figure 3, the accuracy of product restriction III is higher than that of product restriction II. Despite their limitation to a few settings, Figure 3 and Table 2 provides valuable guidance regarding the accuracy versus run-time trade-off for mean field variational Bayes approaches to approximate inference for linear mixed models with crossed random effects.

A more challenging problem is that of meaningful timing comparisons with Markov chain Monte Carlo alternatives to the Algorithms 1–4 streamlined mean field variational Bayes strategies. First, there is the issue that the elapsed computation time for mean field variational Bayes approach is governed by the number of iterations, whereas for Markov chain Monte Carlo approaches it is sample size. Ideally notions of convergence could be used to arrive at comparable stopping rules. But this has its own difficulties due to factors such as tolerance choice and chain stickiness. The accuracy comparisons in Figures 2 and 4 involved the default Markov chain Monte Carlo implementation used by the `rstan` package. For the first three sample size pairs of Table 2 `rstan`, with warm-up and kept sample sizes of 1000, required between 50 and 400 times the computational time compared with mean field variational Bayes with product restriction III and failed to compute for the fourth sample size pair. However, it is well-known that general purpose Bayesian inference engines such as rstan tend to be considerably slower than fit-for-purpose code. We implemented the plain block Gibbs sampling algorithm for model (1) in a low-level language. As expected, this was much faster than rstan with respect to number of draws per second. However, plain block Gibbs sampling exhibited extremely poor mixing for the fixed effects parameters with lag 1 autocorrelation values as high as 0.99. For tests involving warm-up and kept sample sizes of 1000 the *effective* sample size, according to the definition used by the `rstan` package, was as low as 5 for the components of $\boldsymbol{\beta}$. The `rstan` effective sample sizes are much higher, typically by a factor of 10 or more. This chain stickiness problem with plain block Gibbs sampling implies a degradation in the quality of its Bayesian inference which stymies fair timing comparisons. Recent work by Papaspiliopoulos, Roberts, and Zanella (2020) provided a theoretical explanation of the poor performance of plain Gibbs sampling for cross random effects models and proposed a remedy for models similar to (1). This new work may lead to competitive scalable alternatives to this article's streamlined mean field variational Bayes approaches for model (1).

5.3. Conclusions from Comparison Studies

Our first conclusion based on the studies described in this section is that product restriction I should not be used for streamlined variational inference since it is much less accurate than product restriction II without any significant speed and storage advantages. Even though the asymmetry of product restriction II is slightly disconcerting, it is better to bear with it in the interest of having the fixed effects posterior density functions approximated more accurately.

The choice between product restrictions II and III depends on the size of the problem, availability of computing resources and the need for speed in the application at hand. If speed is not important then product restriction III is preferable due to its high inferential accuracy. Product restriction II is a fallback for extremely large problems.

6. Illustration for Data From a Large Longitudinal Education Study

We now provide illustration for data from the National Education Longitudinal Study which was launched in the United States in early 1988. Details of the study are given in Thurgood et al. (2003). The data are publicly available from the U.S. National Center for Education Statistics. Our illustration focuses on students within their last 5 years of secondary education. The data involve longitudinal measurements on 8564 students with each student having his or her academic ability assessed according to 24 items. The full list of items is given in Table S.2 of the supplementary materials and includes, for example, test scores in reading, mathematics and science. All data scores are expressed in percentage form. Other variables such as gender and parental education levels were also recorded.

We did not conduct a full and thorough analysis of these data and avoid exploring matters such as careful variable creation and model selection. Instead, we consider an illustrative Bayesian mixed model with a very large number of crossed random effects.

The model we considered is, for $1 \le i \le 8,442$ and $1 \le i' \le 24$,

$$\begin{aligned}&\boldsymbol{y}_{ii'}|\beta_0,\ldots,\beta_5,u_{i0},u_{i1},u'_{i'0},u'_{i'1},\sigma^2\\&\quad\overset{\text{ind.}}{\sim} N\Big(\beta_0+u_{i0}+u'_{i'0}+(\beta_1+u_{i1}+u'_{i'1})\boldsymbol{x}_{1,ii'}\\&\quad+\beta_2\boldsymbol{x}_{2,ii'}+\cdots+\beta_5\boldsymbol{x}_{5,ii'},\sigma^2\boldsymbol{I}\Big),\\&\begin{bmatrix}u_{i0}\\u_{i1}\end{bmatrix}\Big|\boldsymbol{\Sigma}\overset{\text{ind.}}{\sim}N(\boldsymbol{0},\boldsymbol{\Sigma}),\ \begin{bmatrix}u'_{i'0}\\u'_{i'1}\end{bmatrix}\Big|\boldsymbol{\Sigma}'\overset{\text{ind.}}{\sim}N(\boldsymbol{0},\boldsymbol{\Sigma}')\end{aligned}\qquad(12)$$

where $\boldsymbol{y}_{ii'}$ is the $n_{ii'} \times 1$ vector of scores for the ith student and i'th item. The $n_{ii'} \times 1$ predictor vectors $\boldsymbol{x}_{1,ii'}, \ldots, \boldsymbol{x}_{5,ii'}$ are $n_{ii'} \times 1$ vectors containing measurements for the (i, i')th student/item pair on values of the variables $x_1, \ldots, x_5$ which are defined as follows:

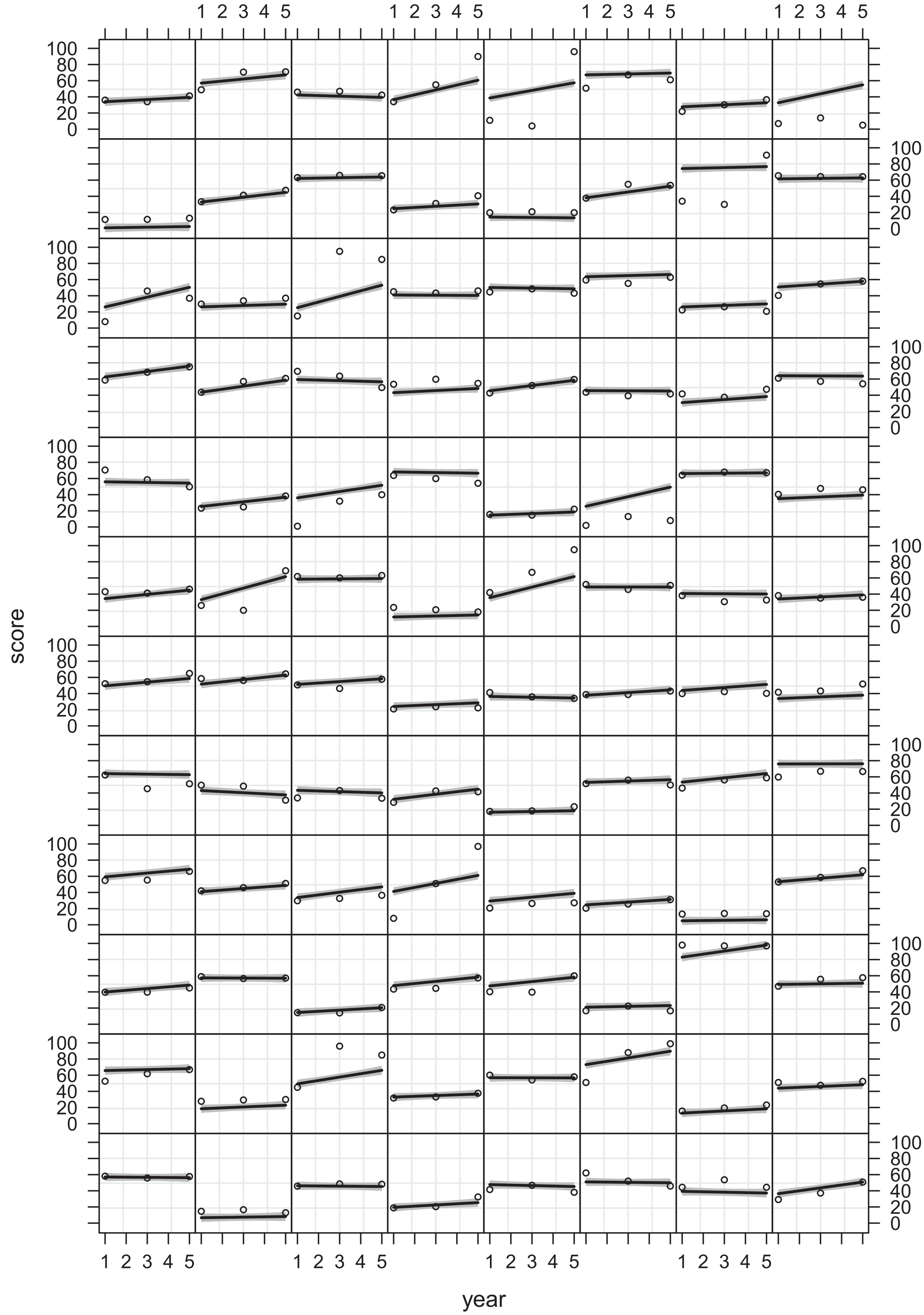

Figure 5. Fitted lines for 96 randomly chosen student/item pairs from the streamlined mean field variational Bayes analysis of data from the National Education Longitudinal Study of 1988. The other predictors are set to their average values. The gray shading corresponds to pointwise 95% credible intervals.

x_1 = year of study (either 1, 3 or 5),

x_2 = indicator that the student is male,

x_3 = indicator that the student spent at least 30 hr per week on homework,

x_4 = indicator that the student's father has at least a high school education, and

x_5 = indicator that the student's mother has at least a high school education.

The priors were set to be

$$\beta_0, \ldots, \beta_5 \overset{\text{ind.}}{\sim} N(0, 10^{10}), \quad \sigma^2 | a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/a_{\sigma^2}),$$
$$a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 10^{-10}),$$

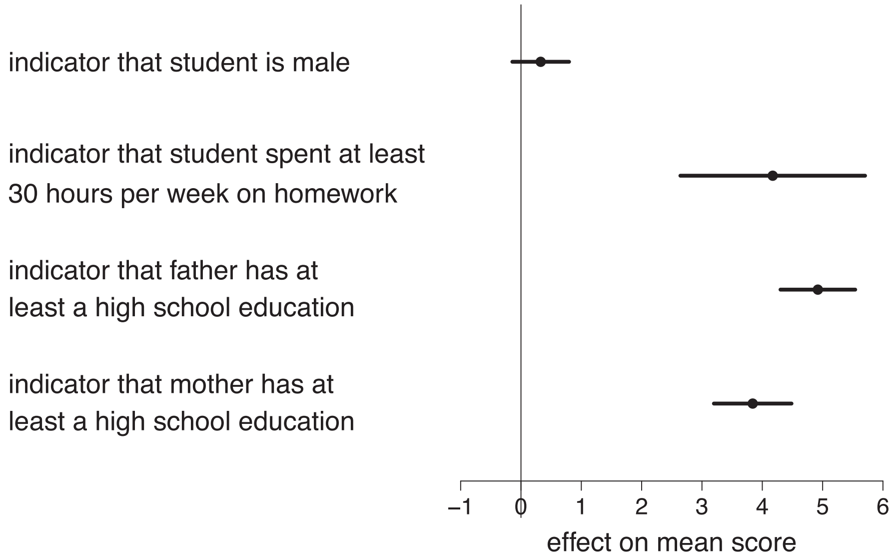

Figure 6. Approximate posterior means (solid dots) and 95% credible intervals (line segments) for $\beta_2, \ldots, \beta_5$ for the mean field variational Bayes, with product restriction III, fit of (12) to data from the National Education Longitudinal Study of 1988.

$$\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, 4, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big),$$
$$\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \frac{2}{10^{10}}\boldsymbol{I}_2),$$
$$\boldsymbol{\Sigma}'|\boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, 4, \boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1}\big) \quad \text{and}$$
$$\boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \frac{2}{10^{10}}\boldsymbol{I}_2).$$

The response data was transformed to the unit interval for Bayesian analysis with these priors. The parameters were then back-transformed to match the original response scale. In addition, to make the Gaussian assumption more plausible, we only considered fields with test scores between 1% and 99% inclusive. We fit model (12) using mean field variational Bayes under product restriction III with Fortran 77 implementation of Algorithm 3 with 100 iterations. Again, we used the third author's MacBook Air laptop with its 2.2 gigahertz processor and 8 gigabytes of random access memory and the fit took just under 5 min.

Figure 5 shows 96 randomly chosen of the random line year effects, corresponding to posterior means, with each of $x_2, \ldots, x_5$ fixed at their average values and the horizontal and vertical ranges set to be the same for each panel. Shading corresponds to pointwise 95% credible intervals. Strong heterogeneity in the year effects across subject/item pairs is apparent, although it should be noted that Figure 5 represents only about 0.05% of all such effects.

Figure 6 provides a graphical summary of the effects of $x_2, \ldots, x_5$. Each line segment corresponds to an approximate 95% credible interval for the corresponding coefficient. The mean field variational Bayes posterior means are shown as solid dots. For example, having a father with at least a high school education leads to an elevation of about 5% in mean test score. The homework and education-related predictors are seen to be highly significant, whereas gender is not significant.

It is apparent from Figure 5 that inclusion of crossed random effects that is crucial for well-grounded estimation and inference concerning the fixed effects, provided by Figure 6. An ordinary least squares analysis with the crossed random effects omitted would ignore the pronounced heterogeneities in the age effects and their subject/item interactions and result in imprecise inference for the Figure 6 effects. Ordinary least squares also ignores within-subject and within-item correlations of the scores, whereas such correlations are accounted for by model (1).

7. Conclusions

We have derived and evaluated three streamlined variational inference algorithms for Gaussian response linear mixed models with crossed random effects, with differing product restriction stringencies. It is concluded that the most stringent algorithm, labeled mean field variational Bayes with product restriction I, should be eliminated from contention which leaves product restriction II and product restriction III. Mean field variational Bayes with product restriction II is shown to be scalable to very large numbers of crossed random effects. Mean field variational Bayes with product restriction III is less scalable but highly accurate. Our numerical results provide valuable guidance for use of our algorithms in terms of accuracy and run-time trade-offs. For moderate problems product restriction III delivers fast and accurate inference. For increasingly large problems, product restriction II offers a scalable alternative.

Supplementary Materials

Supplementary materials for this article are online. Please go to *www.tandfonline.com/r/JCGS*.

Acknowledgments

We are grateful for comments from Doug Bates, Emanuele Degani, and Omar Ghattas.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

This research was partially supported by Australian Research Council Discovery Project DP140100441 and U.S. National Institutes of Health grants R01AA23187, P50DA039838, and U01CA229437.

References

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008), "Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items," *Journal of Memory and Language*, 59, 390–412. [99]

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, New York: Springer. [102,105]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review of Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [100]

Doran, H., Bates, D. M., Bliese, P., and Dowling, M. (2007), "Estimating the Multilevel Rasch Model: With the lme4 Package," *Journal of Statistical Software*, 20, 1–18. [99]

Huang, A., and Wand, M. P. (2013), "Simple Marginally Noninformative Prior Distributions for Covariance Matrices," *Bayesian Analysis*, 8, 439–452. [100,101,102]

Jeon, M., Rijmen, F., and Rabe-Hesketh, S. (2017), "A Variational Maximization-Maximization Algorithm for Generalized Linear Mixed Models with Crossed Random Effects," *Psychometrika*, 3, 693–716. [99,100]

Lee, C. Y. Y., and Wand, M. P. (2016),"Streamlined Mean Field Variational Bayes for Longitudinal and Multilevel Data Analysis," *Biometrical Journal*, 58, 868–895. [103]

Menictas, M., Nolan, T. H., Simpson, D. G., and Wand, M. P. (2021), "Streamlined Variational Inference for Higher Level Group-Specific Curve Models," *Statistical Modelling*, 21, 479–519. [99]

Menictas, M., and Wand, M. P. (2013), "Variational Inference for Marginal Longitudinal Semiparametric Regression," *Stat*, 2, 61–71. [110]

Minka, T., Winn, J., Guiver, Y., Fabian, D., and Bronskill, J. (2018), Infer.NET 0.3, Microsoft Research Cambridge. Available at *http://dotnet.github.io/infer* [100]

Nolan, T. H., Menictas, M., and Wand, M. P. (2020), "Streamlined Computing for Variational Inference with Higher Level Random Effects," *Journal of Machine Learning Research*, 21, 1–62. [99,103,106]

Nolan, T. H., and Wand, M. P. (2017), "Accurate Logistic Variational Message Passing: Algebraic and Numerical Details," *Stat*, 6, 102–112. [100]

——— (2020), "Solutions to 'Multilevel Sparse Matrix Problems," ANZIAM Journal, 62, 18–41. [99,103]

Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153. [102]

Papaspiliopoulos, O., Roberts, G. O., and Zanella, G. (2020), "Efficient Parameterisations for Normal Linear Mixed Models," *Biometrika*, 107, 25–40. [112]

R Core Team (2019), R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria. Available at *https://www.R-project.org/*. [106]

Stan Development Team (2021), RStan: The R Interface to Stan. R package version 2.26.3. Available at *https://mc-stan.org/*. [106]

Thurgood, L., Walter, E., Carter, G., Henn, S., Huang, G., Nooter, D., Smith, W., Cash, R. W., and Salvucci, S. (2003), "National Education Longitudinal Study of 1988 (NELS: 88)," in *NCES Handbook of Survey Methods*, eds. M. Seastrom, T. Phan and M. Cohen, pp. 53-66, Washington D.C.: U.S. Department of Education, National Center for Education Studies. [112]

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016), *Edward: A Library for Probabilistic Modeling, Inference, and Criticism*. Unpublished manuscript. Available at *https://arxiv.org/abs/1610.09787*. [100]

Wand, M. P., and Jones, M. C. (1995), *Kernel Smoothing*, London: Chapman and Hall. [107]

Wand, M. P. (2017), "Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing," (with Discussion), *Journal of the American Statistical Association*, 112, 137–168. [100,106]

Winn, J., and Bishop, C. M. (2005), "Variational Message Passing," *Journal of Machine Learning Research*, 6, 661–694. [100]