Streamlined Variational Inference for Linear Mixed Models with Crossed Random Effects

By Marianne Menictas[1], Gioia Di Credico[2] and Matt P. Wand[3]

Harvard University[1], *Universitá di Padova*[2] *and University of Technology Sydney*[3]

12th December, 2019

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.

Keywords: Mean field variational Bayes; item response theory; 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 *et al.*, 2008; Jeon *et al.*, 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 non-random. 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 data sets.

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 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 & 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.

If variational inference is applied to models containing random effects then a crucial modification of the algebra is matrix algebraic *streamlining*. This is because the random effects design matrices are often sparse and potentially very large. Clever algorithms that recognize the sparseness patterns can lead to dramatic savings in terms of storage and computing time. Nolan *et al.* (2019) 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.* (2019). 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 & Wand, 2019) 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.

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 *et al.* (2017) which applied the notions of Gaussian variational approximation to frequentist generalized linear mixed models with crossed random effects. Jeon *et al.* (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 & 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 paper 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 the 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. An online supplement contains derivational and related details. Some results for frequentist inference for crossed random effects are also given in the online supplement.

2 Bayesian Crossed Random Effects Linear Mixed Models

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

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

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 $y_{ii'}$ 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. 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 & Wand (2013).

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

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

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 $\Lambda_{\boldsymbol{\Sigma}}$ and $\Lambda_{\boldsymbol{\Sigma}'}$. Prior specification (B) involves:

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

for hyperparameters $\nu_{\sigma^2},\nu_{\boldsymbol{\Sigma}},\nu_{\boldsymbol{\Sigma}'},s_{\boldsymbol{\Sigma},1}^2,\ldots,s_{\boldsymbol{\Sigma},q}^2,s_{\boldsymbol{\Sigma}',1}^2,\ldots,s_{\boldsymbol{\Sigma}',q'}^2>0$. As explained in Huang & 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:

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

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.

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

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

as well as

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

Also, we define

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

and

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

Introducing the dimensional notation:

$$n_{i\bullet} \equiv \sum_{i'=1}^{m'} n_{ii'}, \quad 1 \le i \le m, \quad n_{\bullet i'} \equiv \sum_{i=1}^{m} n_{ii'}, \quad 1 \le i' \le m', \qquad \text{and} \quad n_{\bullet\bullet} \equiv \sum_{i=1}^{m}\sum_{i'=1}^{m'} n_{ii'}$$

we then have the dimensions of each of the new data matrices being as follows:

$$\overset{\blacktriangle}{\boldsymbol{y}}_i \text{ is } n_{i\bullet} \times 1, \quad \overset{\blacktriangle}{\boldsymbol{X}}_i \text{ is } n_{i\bullet} \times p, \quad \overset{\blacktriangledown}{\boldsymbol{y}}_{i'} \text{ is } n_{\bullet i'} \times 1, \quad \overset{\blacktriangledown}{\boldsymbol{X}}_{i'} \text{ is } n_{\bullet i'} \times p,$$

$$\overset{\blacktriangle}{\boldsymbol{Z}}_i \text{ is } n_{i\bullet} \times q, \quad \overset{\blacksquare}{\boldsymbol{Z}}{}'_i \text{ is } n_{i\bullet} \times (m'q'), \quad \overset{\blacktriangledown}{\boldsymbol{Z}}{}'_{i'} \text{ is } n_{\bullet i'} \times q',$$

$$\boldsymbol{y} \text{ is } n_{\bullet\bullet} \times 1, \quad \boldsymbol{X} \text{ is } n_{\bullet\bullet} \times p \quad \text{and} \quad \boldsymbol{Z} \text{ is } n_{\bullet\bullet} \times (mq + m'q').$$

3 Variational Inference

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

$$\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}' | \boldsymbol{y}). \tag{5}$$

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

$$\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}') \tag{6}$$

be a mean field approximation of (5). Several product restrictions can be placed on the $\mathfrak{q}$-density function in (6). Here we consider three such restrictions:

$$\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}') = \begin{cases} \mathfrak{q}(\boldsymbol{\beta})\mathfrak{q}(\boldsymbol{u})\mathfrak{q}(\boldsymbol{u}')\,\mathfrak{q}(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{labeled } \textit{product restriction I,} \\ \mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})\mathfrak{q}(\boldsymbol{u}')\,\mathfrak{q}(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{labeled } \textit{product restriction II,} \\ \mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')\,\mathfrak{q}(\sigma^2, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}'), & \text{labeled } \textit{product restriction III.} \end{cases} \tag{7}$$

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 large variances. On the other hand, product restriction III allows for joint posterior covariance matrix of $(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}')$ in its $\mathfrak{q}$-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 (7) conveys the product restrictions in their minimal forms. However, conditional independencies inherent in (1) mean that additional factorizations ensue as follows:

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

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

$$\mathfrak{q}(\sigma^2, a_{\sigma^2}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}'}) = \mathfrak{q}(\sigma^2)\mathfrak{q}(a_{\sigma^2})\mathfrak{q}(\boldsymbol{\Sigma})\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})\mathfrak{q}(\boldsymbol{\Sigma})\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}).$$

Under either product restrictions I, II or III, and letting $\boldsymbol{u}_{\text{all}} \equiv (\boldsymbol{u}, \boldsymbol{u}')$, the usual steps lead to the $\mathfrak{q}$-density functions of the model parameters having the following forms:

$$\begin{aligned} &\mathfrak{q}^*(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}}) \text{ has a } N\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}\big) \text{ distribution,} \\ &\mathfrak{q}^*(\sigma^2) \text{ has an Inverse-}\chi^2\big(\xi_{\mathfrak{q}(\sigma^2)}, \lambda_{\mathfrak{q}(\sigma^2)}\big) \text{ distribution,} \\ &\mathfrak{q}^*(\boldsymbol{\Sigma}) \text{ has an Inverse-G-Wishart}(G_{\text{full}}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}) \text{ distribution} \\ \text{and} \quad &\mathfrak{q}^*(\boldsymbol{\Sigma}') \text{ has an Inverse-G-Wishart}(G_{\text{full}}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}) \text{ distribution.} \end{aligned} \tag{8}$$

The naïve updates for $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ may be written

$$\begin{aligned} &\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})} \leftarrow (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1}(\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{y} + \boldsymbol{o}_{\text{MFVB}}) \quad \text{and} \\ &\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})} \leftarrow (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1} \end{aligned} \tag{9}$$

where $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$,

$$\begin{aligned} &\boldsymbol{R}_{\text{MFVB}} \equiv \mu_{\mathfrak{q}(1/\sigma^2)}^{-1}\boldsymbol{I}, \quad \boldsymbol{D}_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I}_{m'} \otimes \boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})} \end{bmatrix} \\ &\text{and} \quad \boldsymbol{o}_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}. \end{aligned} \tag{10}$$

For variational inference concerning within cell effects as well as σ^2, $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$ the following sub-blocks of $\mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ are required:

$$\begin{array}{l}\mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}, \quad \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}, \quad E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}, \quad 1\le i\le m,\\ \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}, \quad E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}, \quad 1\le i'\le m',\\ E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}, \quad 1\le i\le m,\ 1\le i'\le m'.\end{array} \tag{11}$$

Assuming that m' is moderately sized, the matrices in (11) have reasonable sizes and storage of these is likely to be feasible. Under product restriction III, each of the matrices in (11) potentially is non-zero. However, under product restrictions I and II we have

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}=\boldsymbol{O}, \quad 1\le i'\le m',$$

and

$$E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}=\boldsymbol{O}, \quad 1\le i\le m,\ 1\le i'\le m'.$$

It follows that for product restriction II, the non-zero sub-blocks of $\mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ are

$$\mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})},\ \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)},\ E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}, \quad 1\le i\le m,\ \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})},\ 1\le i'\le m'. \tag{12}$$

Under product restriction I we also have

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}=\boldsymbol{O}, \quad 1\le i\le m.$$

Therefore, product restriction I is such that the non-zero sub-blocks of $\mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ are

$$\mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}, \quad \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)},\ 1\le i\le m, \quad \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})},\ 1\le i'\le m'. \tag{13}$$

4 Streamlined Variational Inference

In this section we provide streamlined variational inference algorithms under product restrictions I, II and III. However, as we will soon explain, in the case of product restriction III we require m' to be moderately sized in order for the label "streamlined" to be valid.

Variational inference for σ^2, $\mathbf{\Sigma}$ and $\mathbf{\Sigma}'$ 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 $\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})$ so these are treated separately in each of the next three subsections. 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.2 of the online supplement.

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

$$\begin{array}{l}\boldsymbol{b}\longleftarrow\left[\begin{array}{c}\mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\left[\boldsymbol{y}-\underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}\Big(\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}+\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\Big)\right\}\right]\\ \mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}}\end{array}\right]\\ \boldsymbol{B}\longleftarrow\left[\begin{array}{c}\mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{X}\\ \mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\end{array}\right] \quad ; \quad \mathcal{S}\longleftarrow \textsc{SolveLeastSquares}\Big(\{\boldsymbol{b},\boldsymbol{B}\}\Big)\\ \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow \boldsymbol{x}\text{ component of }\mathcal{S} \quad ; \quad \mathbf{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow(\boldsymbol{B}^T\boldsymbol{B})^{-1}\text{ component of }\mathcal{S}\end{array} \tag{14}$$

where the SOLVELEASTSQUARES algorithm is given by Algorithm S.1 in Section S.3 of the online supplement. Then, given the current values of the $\mathfrak{q}$-density parameters of $\boldsymbol{\beta}$, $\boldsymbol{u}'$, σ^2 and $\boldsymbol{\Sigma}$ the updates for the parameters of the $\mathfrak{q}(\boldsymbol{u}_i)$, $1 \le i \le m$, have similar expressions involving the SOLVELEASTSQUARES algorithm. The updates for $\mathfrak{q}(\boldsymbol{u}'_{i'})$, $1 \le i' \le m'$, are analogous.

The full set of updates is provided by Algorithm 1.

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

$\mathfrak{q}$-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.

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

$$\boldsymbol{B} \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathfrak{q}(1/\sigma^2)} \boldsymbol{X} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \end{bmatrix} \quad ; \quad \mathcal{S} \longleftarrow \text{SOLVELEASTSQUARES}\Big(\{\boldsymbol{b}, \boldsymbol{B}\} \Big)$$

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

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

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

$$\boldsymbol{B} \longleftarrow \begin{bmatrix} \mu^{1/2}_{\mathfrak{q}(1/\sigma^2)} \overset{\blacktriangle}{\boldsymbol{Z}}_i \\ \boldsymbol{M}^{1/2}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} \end{bmatrix} \quad ; \quad \mathcal{S} \longleftarrow \text{SOLVELEASTSQUARES}\Big(\{\boldsymbol{b}, \boldsymbol{B}\} \Big)$$

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

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

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

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

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

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

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

Under product restriction II the updates for the $\mathfrak{q}(\boldsymbol{u}'_{i'})$ parameters are the same as those for product restriction I. However streamlined updating of the $\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})$ parameters is more delicate. The problem can be embedded within the class of two-level sparse matrix problems as defined in Nolan & Wand (2019) and is encapsulated in Result 1. Note that Result 1 uses matrix sub-block notation given by (S.2) in Section S.4 of the online supplement. The derivation of this result is given in Section S.5 of the online supplement of this article.

Result 1. *According to product restriction II, the mean field variational Bayes updates of* $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ *and each of the sub-blocks of* $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ *listed in (12), given the current values of* $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}$, $1 \le i' \le m'$, *are expressible as a two-level sparse matrix least squares problem of the form:*

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

where $\boldsymbol{b}$ *and the non-zero sub-blocks of* $\boldsymbol{B}$, *according to the notation in (S.1) of the online supplement, are, for* $1 \le i \le m$,

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

and

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

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

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

and

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

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) in the online supplement.*

Result 1 gives rise to Algorithm 2, which provides the full set of updates of the $\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters under product restriction II. Note that Algorithm 2 makes use of the SolveTwoLevelSparseLeastSquares algorithm from Nolan *et al.* (2019) and reproduced for convenience in Section S.4 of the online supplement.

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 & Wand (2016) and Nolan *et al.* (2019).

Result 2 embeds the updates of the $\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ parameters within the class of two-level sparse matrix problems as defined in Nolan & Wand (2019) and summarized in Section S.4 of the online supplement. 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.

Algorithm 2 *Mean field variational Bayes algorithm for updating the parameters of* $\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ *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: $\boldsymbol{\mu_\beta}(p \times 1)$, $\boldsymbol{\Sigma_\beta}(p \times p)$ symmetric and positive definite.

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

$\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}(q' \times q')$ both symmetric and positive definite.

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

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

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

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

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

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

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

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

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

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

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

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

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

$\Big\{ E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T :\ 1 \le i \le m \Big\}$

Result 2. *According to product restriction III, the mean field variational Bayes updates of* $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})}$ *and each of the sub-blocks of* $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})}$ *in (11) is expressible as a two-level sparse matrix least squares problem of the form:*

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

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

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

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

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),

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} = \textit{first } p \textit{ rows of } \boldsymbol{x}_1, \quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} = \textit{ top left } p \times p \textit{ sub-block of } \boldsymbol{A}^{11},$$

$$\underset{1 \le i' \le m'}{\text{stack}}(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}) = \textit{subsequent } (m'q') \times 1 \textit{ entries entries of } \boldsymbol{x}_1 \textit{ following } \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})},$$

$$E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})})^T\} = \textit{ subsequent } p \times q' \textit{ sub-blocks of } \boldsymbol{A}^{11} \textit{ to the right of } \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})},$$

$$\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}'_{i'})} = \textit{ subsequent } q' \times q' \textit{ diagonal sub-blocks of } \boldsymbol{A}^{11} \textit{ following } \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}, \quad 1 \le i' \le m',$$

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} = \boldsymbol{x}_{2,i}, \quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} = \boldsymbol{A}^{22,i}, \quad E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} = \textit{first } p \textit{ rows of } \boldsymbol{A}^{12,i}$$

$$\textit{and } \underset{1 \le i' \le m'}{\text{stack}}\Big([E_{\mathsf{q}}\{(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})})^T\}]^T\Big) = \textit{ remaining } m'q' \textit{ rows of } \boldsymbol{A}^{12,i}, \quad 1 \le i \le m,$$

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) in the online supplement.

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) but with the $\boldsymbol{B}_i$ and $\overset{\bullet}{\boldsymbol{B}}_i$ sub-blocks specific to Result 2. The yellow regions correspond to the two-level sparsity due to the block diagonal positioning of the $\overset{\bullet}{\boldsymbol{B}}_i$, $1 \le i \le$m. The orange 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 blue/orangeblock on the left is small relative to the remainder of the matrix. The SolveTwoLevelSparseLeastSquares algorithm, listed as Algorithm S.2 in Section S.4 of the online supplement, 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 orange 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 yellow 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

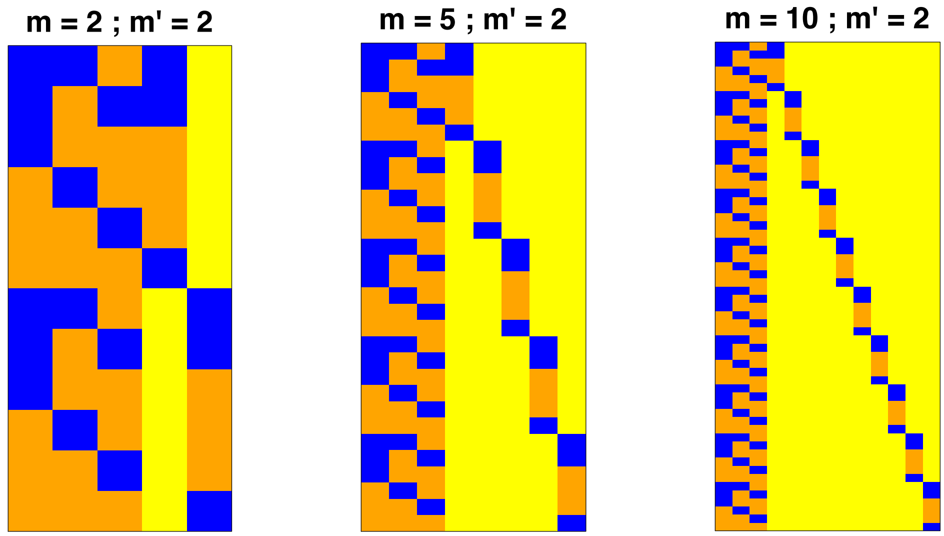

Figure 1: *Image plot representation of the two-level sparse matrix $\boldsymbol{B}$ with generic form given by (S.1) and with sub-blocks as defined in Result 2. The dimension variables are $p = q = q' = n_{ii'} = 1$, $m \in \{2, 5, 10\}$ and $m' = 2$. Blue indicates non-zero entries of $\boldsymbol{B}$.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

given by (3) then standard mean field variational Bayes arguments (e.g. Bishop, 2006; Sections 10.1–10.3) lead to $\xi_{\mathfrak{q}(\sigma^2)} = \xi_{\sigma^2} + n_{\bullet\bullet}$ and

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

Under product restriction I the trace term reduces to

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

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

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

Algorithm 3 *Mean field variational Bayes algorithm for updating the parameters of* $\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\textit{all}})$ *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,

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

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

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

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

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

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

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

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

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

$i_{\text{end}} \longleftarrow i_{\text{stt}} + q' - 1$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})} \longleftarrow$ sub-vector of $\boldsymbol{x}_1$ component of $\mathcal{S}$ with entries i_{stt} to i_{end}

$\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})} \longleftarrow$ 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}

$E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\} \longleftarrow$ sub-block of $\boldsymbol{A}^{11}$ component of $\mathcal{S}$ with rows 1 to p and columns i_{stt} to i_{end}

$i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$

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

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

$E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\} \longleftarrow$ 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 $\mathcal{S}$; $i_{\text{stt}} \longleftarrow p + 1$

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

$i_{\text{end}} \longleftarrow i_{\text{stt}} + q' - 1$

$E_q\{(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\} \longleftarrow$ sub-matrix of $\boldsymbol{\Omega}^T$ with columns i_{stt} to i_{end}

$i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$

continued on a subsequent page ...

Algorithm 3 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

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

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

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

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)

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

If priors (4)

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

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

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

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

Cycle:

If product restriction I then: call Algorithm 1 to update $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ and quantities in (13).
If product restriction II then: call Algorithm 2 to update $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ and quantities in (12).
If product restriction III then: call Algorithm 3 to update $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})}$ and quantities in (11).
If priors (A): $\lambda_{\mathsf{q}(\sigma^2)}\longleftarrow\lambda_{\sigma^2}\ \ ;\ \ \Lambda_{\mathsf{q}(\boldsymbol{\Sigma})}\longleftarrow\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}}\ \ ;\ \ \Lambda_{\mathsf{q}(\boldsymbol{\Sigma}')}\longleftarrow\boldsymbol{\Lambda}_{\boldsymbol{\Sigma}'}$
If priors (B): $\lambda_{\mathsf{q}(\sigma^2)}\longleftarrow\mu_{\mathsf{q}(1/a_{\sigma^2})}\ \ ;\ \ \Lambda_{\mathsf{q}(\boldsymbol{\Sigma})}\longleftarrow\boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\ \ ;\ \ \Lambda_{\mathsf{q}(\boldsymbol{\Sigma}')}\longleftarrow\boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1})}$
For $i=1,\ldots,m$:

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

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

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

If product restriction II or III:

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

If product restriction III:

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

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

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 $i = 1, \ldots, m$:

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

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

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

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

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

If priors (B):

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

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

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

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

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

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

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

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

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

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

$$\boldsymbol{\beta}_{\text{true}} = \begin{bmatrix} 0.58 \\ 1.89 \end{bmatrix}, \quad \sigma^2_{\text{true}} = 0.3, \quad \boldsymbol{\Sigma}_{\text{true}} = \begin{bmatrix} 0.46 & -0.19 \\ -0.19 & 0.17 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}'_{\text{true}} = \begin{bmatrix} 0.3 & -0.12 \\ -0.12 & 0.25 \end{bmatrix}. \quad (15)$$

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

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, 2019) 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 estimation, with direct plug-in bandwidth selection (e.g. Wand & Jones, 1995; Section 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. 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.

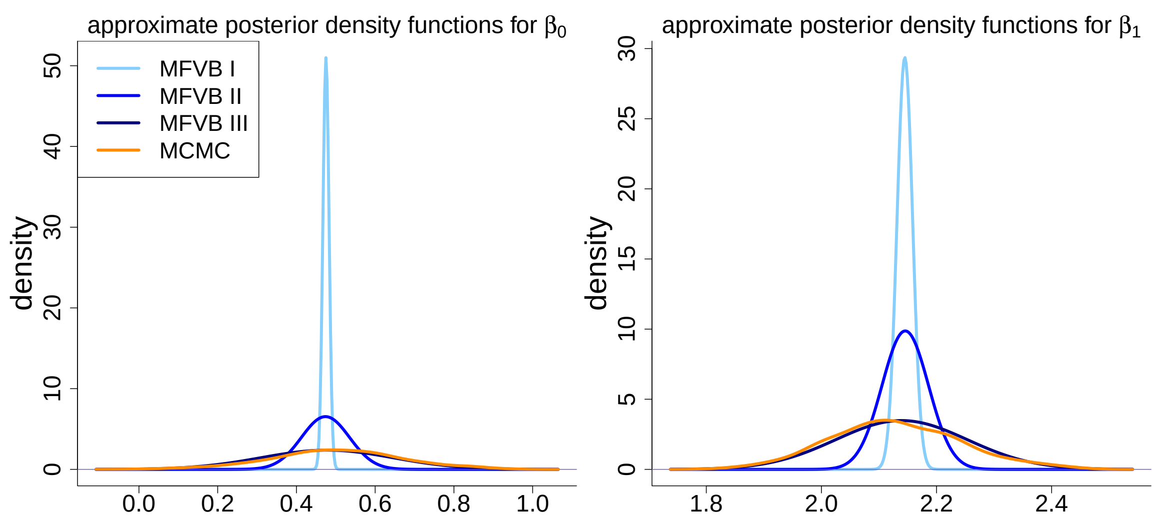

Figure 2: *Approximate posterior density functions for β_0 and β_1, according to three different mean field variational Bayes(MFVB) schemes and Markov chain Monte Carlo (MCMC), for the first replication of the simulation study. The legend uses the abbreviation "MFVB I" for the mean field variational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

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:

$$\text{accuracy} \equiv 100\left\{1 - \tfrac{1}{2}\int_{-\infty}^{\infty} \big|\mathfrak{q}(\theta) - \mathfrak{p}(\theta|\boldsymbol{y})\big|\, d\theta\right\}\%. \tag{16}$$

Note that $0\% \le \text{accuracy} \le 100\%$ with a score of 100% if $\mathfrak{q}(\theta)$ 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 $\mathfrak{q}(\theta)$ is available 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 $\mathbf{\Sigma}$:

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

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

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

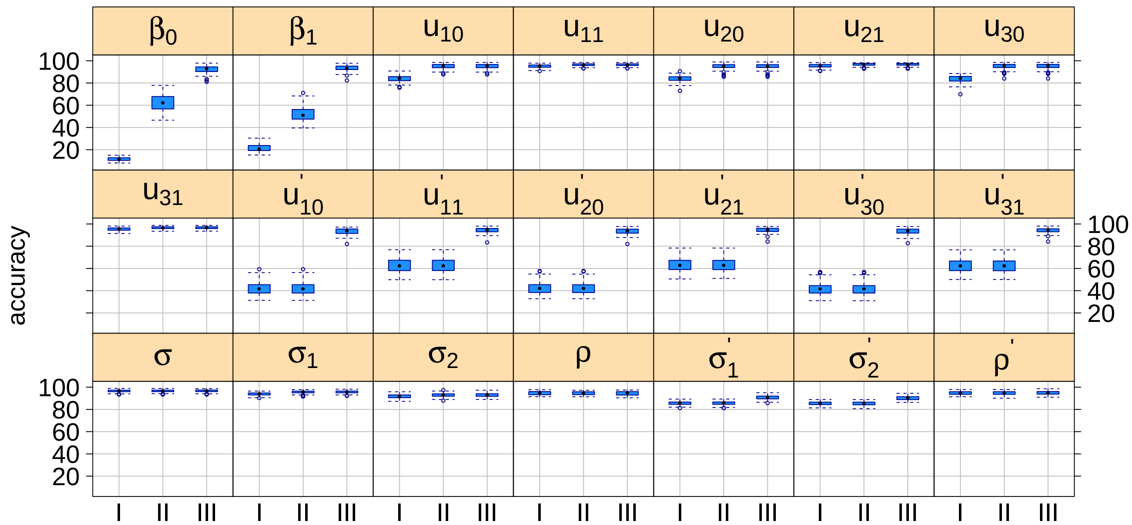

Figure 3: *Side-by-side boxplots for the accuracy scores for 21 parameters and random effects from the simulation study, with accuracy defined according to (16). Each panel corresponds to a separate parameter or random effect and contains side-by-side boxplots for product restrictions I, II and III.*

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 data set, is always 92% or higher. The boxplots in Figure 3 indicate that excellent accuracy is typical for this particular simulation setting.

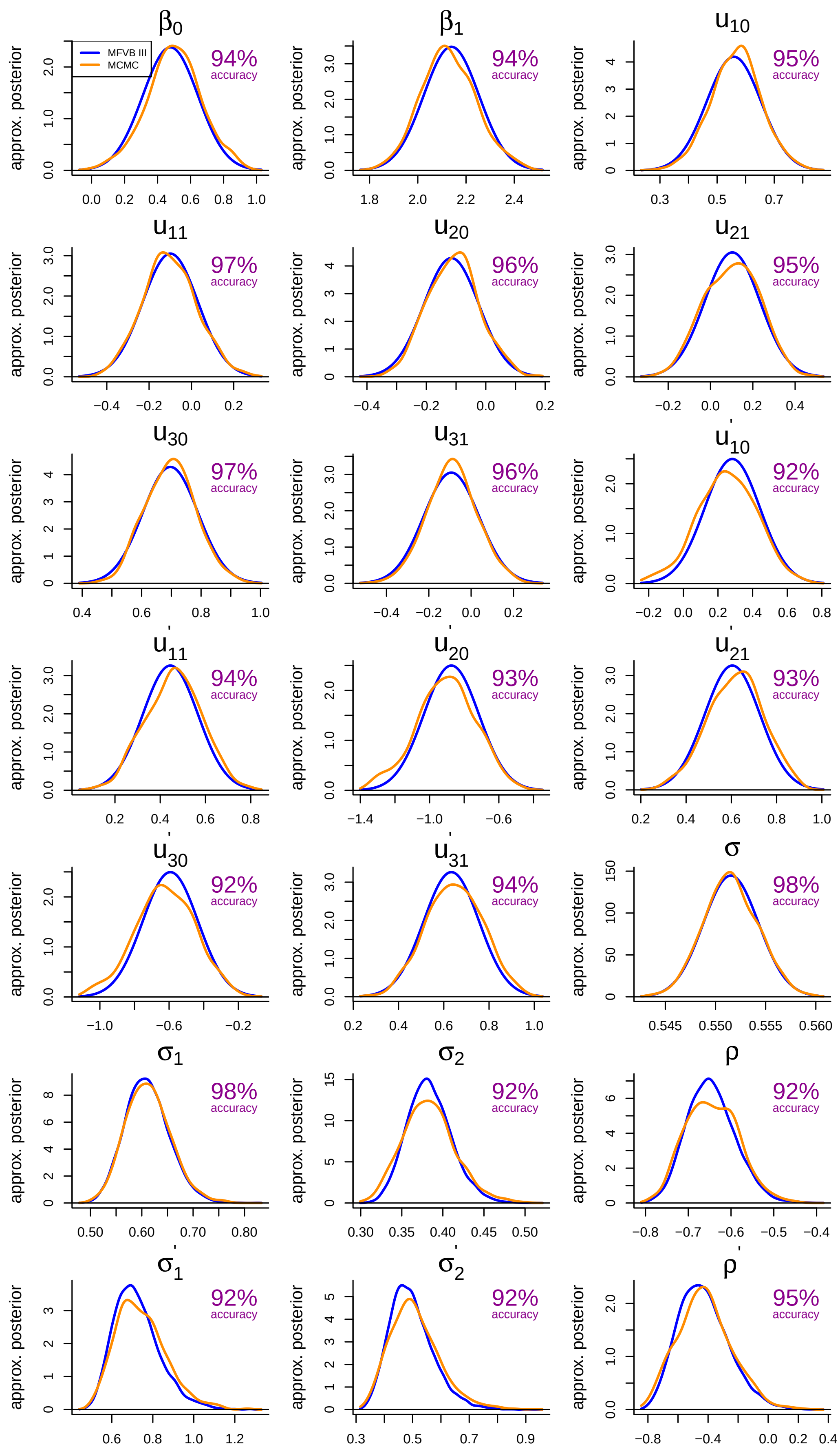

Figure 4: *Approximate posterior density functions for the 21 parameters and random effects for the first replication of the simulation study. The blue curves are posterior density functions obtained using mean field variational Bayes with product restriction III and the orange curves are based on Markov chain Monte Carlo. The accuracy percentages are defined according to (16).*

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 subsection's simulation study – but with increasing crossed random effects dimensions. Specifically, the data were generated according to (15) 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 restriction II,
- mean field variational Bayes with product restriction III,
- Markov chain Monte Carlo.

The mean field variational Bayes computations were performed using Algorithm 4, with calls to Algorithms 2, 3, S.1 and S.2. All five algorithms were implemented in the fast `Fortran 77` language. The number of mean field variational Bayes iterations was fixed at 100. The Markov chain Monte Carlo computations were performed via the package rstan (Stan Development Team, 2019) with 1000 warm-up samples and 1000 retained samples. 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.

Before presenting the results it is admitted up front that speed studies such as this come with caveats such as the effect of the programming language, Markov chain Monte Carlo sample sizes and number of variational inference iterations. Nevertheless, they provide important guidance with regards to scalability to very large problems.

Table 1 lists the average and standard deviation times in seconds. Ratios between the average times for the two slower methods compared with the fastest method (mean field variational Bayes with product restriction II) are also given. In that case of $(m, m') = (800, 160)$, Markov chain Monte Carlo failed to run due to its memory requirements not being met.

(m, m')	MFVB II	MFVB III	MCMC	$\frac{\text{MFVB III}}{\text{MFVB II}}$	$\frac{\text{MCMC}}{\text{MFVB II}}$
(100,20)	0.267 (0.0267)	4.93 (0.082)	1840 (404.1)	18.4	6880
(200,40)	1.44 (0.0996)	66.8 (1.40)	10500 (1370)	46.5	7300
(400,80)	8.92 (0.587)	1130 (8.48)	56900 (13200)	127	6380
(800,160)	54.7 (1.54)	21300 (41.0)	NA (NA)	390	NA

Table 1: *Summaries of results for the speed assessment and comparison study. First three columns: Average (standard deviation) time in seconds for each method, where "MCMC" is short for "Markov chain Monte Carlo", "MFVB II" is short for the mean field variational Bayes according to product restriction II and "MFVB III" is defined similarly. Last two columns: ratios of average times as indicated by the column heading. The NA entries are indicative of the MCMC approach not being feasible for* $(m, m') = (800, 160)$.

Table 1 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 seconds required for the second largest $(m, m') = (400, 80)$ situation. In contrast, Markov chain Monte Carlo could not handle the largest case and, on average, took close to 16 hours for the second case. Across all settings Markov chain Monte Carlo is seen to be more than six thousand times slower than mean field variational Bayes with product restriction II. 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 hours 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 1 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.

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 8,564 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.1 of the online supplement 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 $1 \le i' \le 24$,

$$\begin{aligned}&\boldsymbol{y}_{ii'}|\beta_0,\ldots,\beta_5,u_{i0},u_{i1},u'_{i'0},u'_{i'1},\sigma^2 \stackrel{\text{ind.}}{\sim} N\Big(\beta_0+u_{i0}+{u'}_{i'0}+(\beta_1+u_{i1}+{u'}_{i'1})\boldsymbol{x}_{1,ii'}\\&\quad+\beta_2\boldsymbol{x}_{2,ii'}+\ldots+\beta_5\boldsymbol{x}_{5,ii'},\sigma^2\boldsymbol{I}\Big),\ \left[\begin{array}{c}u_{i0}\\u_{i1}\end{array}\right]\Big|\boldsymbol{\Sigma}\stackrel{\text{ind.}}{\sim}N(\boldsymbol{0},\boldsymbol{\Sigma}),\ \left[\begin{array}{c}{u'}_{i'0}\\{u'}_{i'1}\end{array}\right]\Big|\boldsymbol{\Sigma}'\stackrel{\text{ind.}}{\sim}N(\boldsymbol{0},\boldsymbol{\Sigma}')\end{aligned} \tag{17}$$

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:

$x_1 =$ 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 hours 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

$$\beta_0, \ldots, \beta_5 \overset{\text{ind.}}{\sim} N(0, 10^{10}), \quad \sigma^2 | a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/a_{\sigma^2}), \quad a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 10^{-10}),$$
$$\boldsymbol{\Sigma} | \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, 4, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big), \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \tfrac{2}{10^{10}} \boldsymbol{I}_2),$$
$$\boldsymbol{\Sigma}' | \boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, 4, \boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1}\big) \text{ and } \boldsymbol{A}_{\boldsymbol{\Sigma}'} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \tfrac{2}{10^{10}} \boldsymbol{I}_2).$$

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 considered fitting model (17) using:

(1) Markov chain Monte Carlo via the rstan (Stan Development Team, 2019) R package with a warmup of size 1000 and retained samples of size 1000, and

(2) 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. The attempt at fitting (17) to the National Education Longitudinal Study via rstan was such that rstan ran for more than three days but then failed to save the output, presumably due to the size of the model and data set. The mean field variational Bayes algorithm ran successfully and took just under 5 minutes.

We now present some graphical summaries of the mean field variational Bayes fits. 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.

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

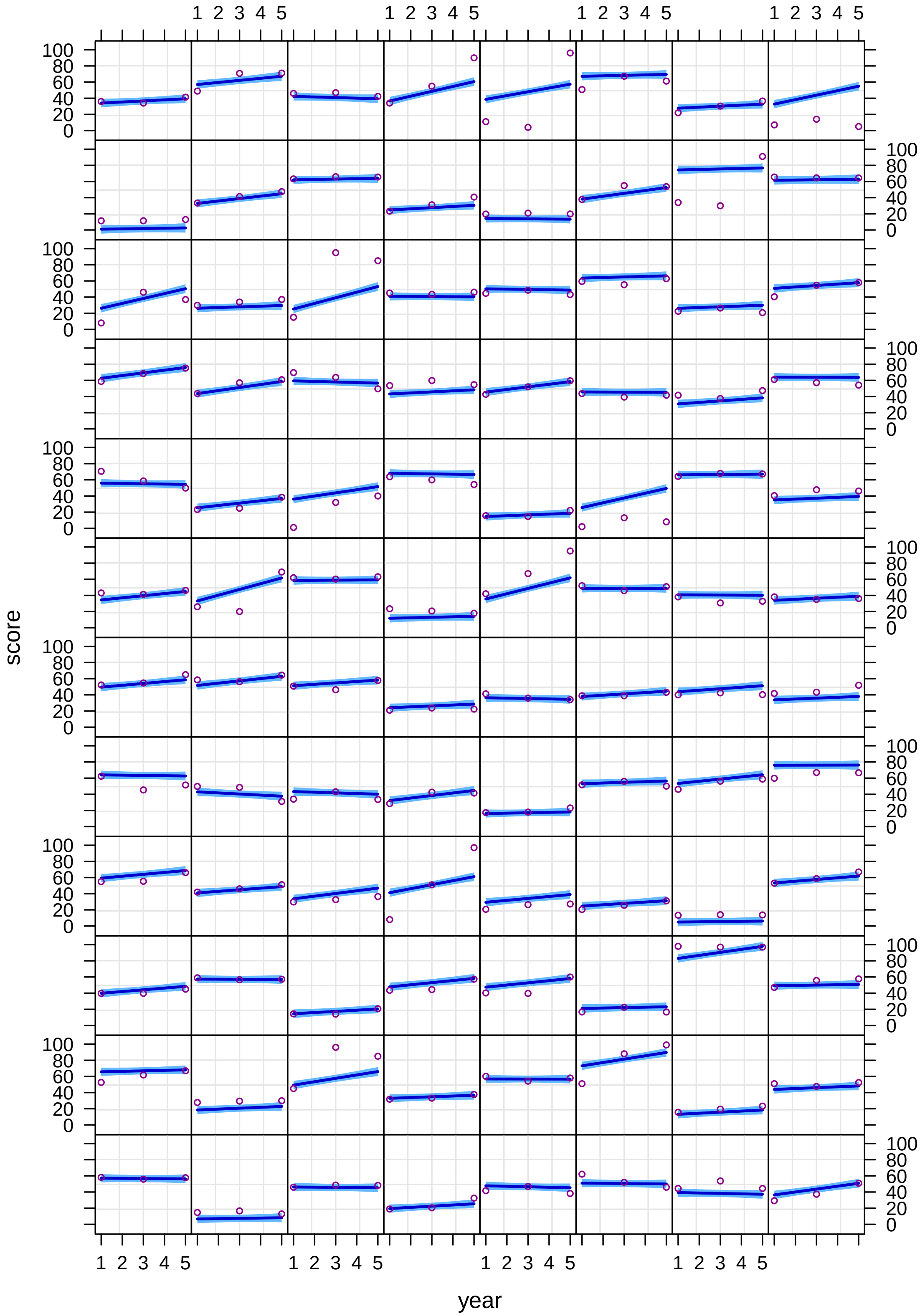

Figure 5: *Fitted lines for 96 randomly chosen student/item pairs from the streamlined mean field variational Bayes analysis of data from the National Education Longitudinal Study of 1988. The other predictors are set to their average values. The light blue shading corresponds to pointwise 95% credible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

with product restriction II is shown to be scalable to very large numbers of crossed random

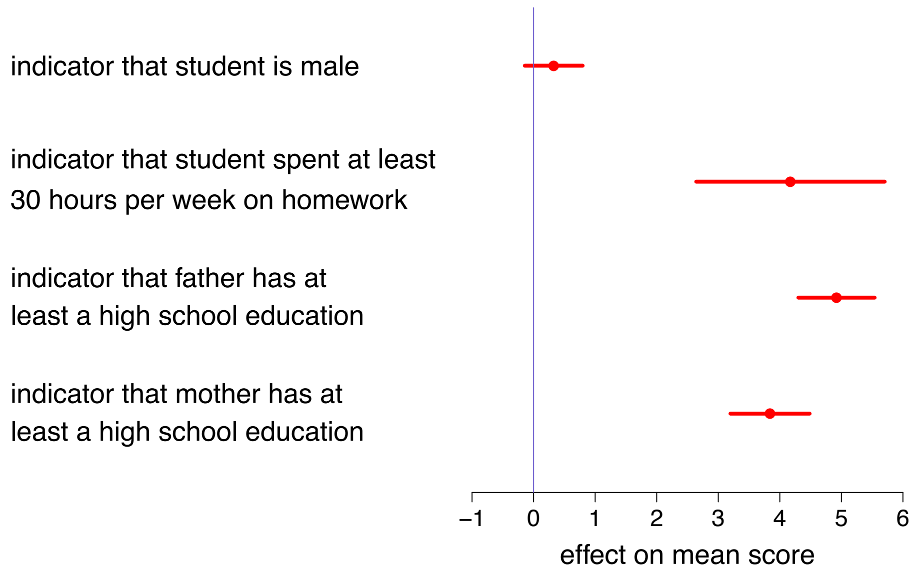

Figure 6: *Approximate posterior means (solid dots) and 95% credible intervals (line segments) for* $\beta_2, \ldots, \beta_5$ *for the mean field variational Bayes, with product restriction III, fit of (17) to data from the National Education Longitudinal Study of 1988.*

effects. Mean field variational Bayes with product restriction III is less scalable but highly accurate. Both approaches are much faster than Markov chain Monte Carlo, which scales very poorly. 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.

Acknowledgements

This research was partially supported by Australian Research Council Discovery Project DP140100441 and U.S. National Institutes of Health grants R01AA23187, P50DA039838 and U01CA229437. We are grateful for comments from Doug Bates and Omar Ghattas.

References

Atay-Kayis, A. & Massam, H. (2005). A Monte Carlo method for computing marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, **92**, 317–335.

Baayen, R.H., Davidson, D.J. & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, **59**, 390–412.

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

Blei, D.M., Kucukelbir, A. & McAuliffe, J.D. (2017). Variational inference: a review of statisticians. *Journal of the American Statistical Association*, **112**, 859–877.

Doran, H., Bates, D.M., Bliese, P. & Dowling, M. (2007). Estimating the multilevel Rasch model: with the lme4 package. *Journal of Statistical Software*, **20**, Issue 2, 1–18.

Jeon, M., Rijmen, F. & Rabe-Hesketh, S. (2017). A variational maximization-maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, **3**, 693–716.

Huang, A. & Wand, M.P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, **8**, 439–452.

Lee, C.Y.Y. & Wand, M.P. (2016). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal*, **58**, 868–895.

Menictas, M., Nolan, T.H., Simpson, D.G. & Wand, M.P. (2019). Streamlined variational inference for higher level group-specific curve models. Unpublished manuscript available at *https://arxiv.org/abs/1903.04043*.

Minka, T., Winn, J., Guiver, Y., Fabian, D. & Bronskill, J. (2018). Infer.NET 0.3, Microsoft Research Cambridge, `http://dotnet.github.io/infer`

Nolan, T.H. & Wand, M.P. (2017). Accurate logistic variational message passing: algebraic and numerical details. *Stat*, **6**,

Nolan, T.H., Menictas, M. & Wand, M.P. (2019). Streamlined computing for variational inference with higher level random effects. Unpublished manuscript available at *https://arxiv.org/abs/1903.06616*.

Nolan, T.H. & Wand, M.P. (2019). Solutions to sparse multilevel matrix problems. Unpublished manuscript available at *https://arxiv.org/abs/1903.03089*.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.18.2. `https://mc-stan.org/`.

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

Tran, D., Kucukelbir, A., Dieng, A.B., Rudolph, M., Liang, D. & Blei, D.M. (2016). *Edward: A library for probabilistic modeling, inference, and criticism.* Unpublished manuscript available at *https://arxiv.org/abs/1610.09787*.

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

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.

Winn, J. & Bishop, C.M. (2005). Variational message passing. *Journal of Machine Learning Research*, **6**, 661–694.

Online Supplement for:

Streamlined Variational Inference for Linear Mixed Models with Crossed Random Effects

By Marianne Menictas[1], Gioia Di Credico[2] and Matt P. Wand[3]

Harvard University[1], *Universitá di Padova*[2] *and University of Technology Sydney*[3]

S.1 The Inverse G-Wishart and Inverse χ^2 Distributions

The Inverse G-Wishart corresponds to the matrix inverses of random matrices that have a *G-Wishart* distribution (e.g. Atay-Kayis & Massam, 2005). For any positive integer d, let G be an undirected graph with d nodes labeled $1, \ldots, d$ and set E consisting of sets of pairs of nodes that are connected by an edge. We say that the symmetric $d \times d$ matrix $\boldsymbol{M}$ *respects* G if

$$\boldsymbol{M}_{ij} = 0 \quad \text{for all} \quad \{i, j\} \notin E.$$

A $d \times d$ random matrix $\boldsymbol{X}$ has an Inverse G-Wishart distribution with graph G and parameters $\xi > 0$ and symmetric $d \times d$ matrix $\boldsymbol{\Lambda}$, written

$$\boldsymbol{X} \sim \text{Inverse-G-Wishart}(G, \xi, \boldsymbol{\Lambda})$$

if and only if the density function of $\boldsymbol{X}$ satisfies

$$\mathfrak{p}(\boldsymbol{X}) \propto |\boldsymbol{X}|^{-(\xi+2)/2} \exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}\,\boldsymbol{X}^{-1})\}$$

over arguments $\boldsymbol{X}$ such that $\boldsymbol{X}$ is symmetric and positive definite and $\boldsymbol{X}^{-1}$ respects G. Two important special cases are

$$G = G_{\text{full}} \equiv \text{totally connected } d\text{-node graph},$$

for which the Inverse G-Wishart distribution coincides with the ordinary Inverse Wishart distribution, and

$$G = G_{\text{diag}} \equiv \text{totally disconnected } d\text{-node graph},$$

for which the Inverse G-Wishart distribution coincides with a product of independent Inverse Chi-Squared random variables. The subscripts of G_{full} and G_{diag} reflect the fact that $\boldsymbol{X}^{-1}$ is a full matrix and $\boldsymbol{X}^{-1}$ is a diagonal matrix in each special case.

The $G = G_{\text{full}}$ case corresponds to the ordinary Inverse Wishart distribution. However, with modularity in mind, we will work with the more general Inverse G-Wishart family throughout this article.

In the $d = 1$ special case the graph $G = G_{\text{full}} = G_{\text{diag}}$ and the Inverse G-Wishart distribution reduces to the Inverse Chi-Squared distributions. We write

$$x \sim \text{Inverse-}\chi^2(\xi, \lambda)$$

for this Inverse-G-Wishart$(G_{\text{diag}}, \xi, \lambda)$ special case with $d = 1$ and $\lambda > 0$ scalar.

S.2 Derivation of the $\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}_{\text{all}})$ Parameters Updates Under Product Restriction I

The full conditional distribution of $\boldsymbol{\beta}$ is

$$\mathfrak{p}(\boldsymbol{\beta}|\text{rest}) \propto \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2)\mathfrak{p}(\boldsymbol{\beta}).$$

Note that $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2)$ can be expressed as the $N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$ density function in the vector

$$\boldsymbol{y} - \underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}\left(\boldsymbol{Z}_{ii'}\boldsymbol{u}_i + \boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'}\right)\right\}$$

Also, $\mathfrak{p}(\boldsymbol{\beta})$ is the $N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$ density function in the vector $\boldsymbol{\beta}$. Then, under product restriction I, standard quadratic form manipulations lead to the optimal $\mathfrak{q}$-density function of $\boldsymbol{\beta}$ being that of the $N(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})})$ distribution with updates

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow (\mu_{\mathfrak{q}(\sigma^2)}\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}(\mu_{\mathfrak{q}(\sigma^2)}\boldsymbol{X}^T\boldsymbol{r}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}^{-1}) \text{ and } \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow (\mu_{\mathfrak{q}(\sigma^2)}\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}.$$

where

$$\boldsymbol{r} \equiv \boldsymbol{y} - \underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le i'\le m'}{\text{stack}}\left(\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)} + \boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\right)\right\}.$$

If $\boldsymbol{b}$ and $\boldsymbol{B}$ are defined according to the updates in (14) then simple algebra shows that

$$\boldsymbol{B}^T\boldsymbol{b} = \mu_{\mathfrak{q}(\sigma^2)}\boldsymbol{X}^T\boldsymbol{r} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}^{-1} \quad \text{and} \quad \boldsymbol{B}^T\boldsymbol{B} = \mu_{\mathfrak{q}(\sigma^2)}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}.$$

Therefore, the $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}$ update corresponds to the least squares solution $\boldsymbol{x} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{b}$ and the update of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}$ corresponds to $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$.

Analogous arguments can be used to justify the updates for the parameters of $\mathfrak{q}(\boldsymbol{u}_i)$, $1 \le i \le m$, and $\mathfrak{q}(\boldsymbol{u}'_{i'})$, $1 \le i' \le m'$.

S.3 The SolveLeastSquares Algorithm

The SolveLeastSquares is concerned with solving the least squares problem

$$\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x}\|^2.$$

which has solution $\boldsymbol{x} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{b}$. The matrix $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$ is also of intrinsic interest. In next subsection a version of this problem is solved for the situation where $\boldsymbol{B}$ has two-level sparse structure. In this subsection there is no sparseness structure imposed on $\boldsymbol{B}$.

Algorithm S.1 SolveLeastSquares *for solving the least squares problem: minimise* $\|\boldsymbol{b} - \boldsymbol{B}\,\boldsymbol{x}\|^2$ *in* $\boldsymbol{x}$ *and obtaining* $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$.

Inputs: $\left\{\boldsymbol{b}(\tilde{n}\times 1),\ \boldsymbol{B}(\tilde{n}\times p)\right\}$

Decompose $\boldsymbol{B} = \boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix}$ such that $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$ and $\boldsymbol{R}$ is upper-triangular.

$\boldsymbol{c} \longleftarrow \boldsymbol{Q}^T\boldsymbol{b}$; $\boldsymbol{c}_1 \longleftarrow$ first p rows of $\boldsymbol{c}$

$\boldsymbol{x} \longleftarrow \boldsymbol{R}^{-1}\boldsymbol{c}_1$; $(\boldsymbol{B}^T\boldsymbol{B})^{-1} \longleftarrow \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$

Output: $\left(\boldsymbol{x}, (\boldsymbol{B}^T\boldsymbol{B})^{-1}\right)$

S.4 The SolveTwoLevelSparseLeastSquares Algorithm

The SolveTwoLevelSparseLeastSquares algorithm solves a sparse version of the the least squares problem:

$$\min_{\boldsymbol{x}} \|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x}\|^2$$

which has solution $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{B}^T\boldsymbol{b}$ where $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ where $\boldsymbol{B}$ and $\boldsymbol{b}$ have the following structure:

$$\boldsymbol{B} \equiv \left[\begin{array}{c|c|c|c|c} \boldsymbol{B}_1 & \overset{\bullet}{\boldsymbol{B}}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \hline \boldsymbol{B}_2 & \boldsymbol{O} & \overset{\bullet}{\boldsymbol{B}}_2 & \cdots & \boldsymbol{O} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{B}_m & \boldsymbol{O} & \boldsymbol{O} & \cdots & \overset{\bullet}{\boldsymbol{B}}_m \end{array}\right] \quad \text{and} \quad \boldsymbol{b} = \left[\begin{array}{c} \boldsymbol{b}_1 \\ \hline \boldsymbol{b}_2 \\ \hline \vdots \\ \hline \boldsymbol{b}_m \end{array}\right]. \tag{S.1}$$

The sub-vectors of $\boldsymbol{x}$ and the sub-matrices of $\boldsymbol{A}$ corresponding to its non-zero blocks of are labeled as follows:

$$\boldsymbol{x} = \left[\begin{array}{c} \boldsymbol{x}_1 \\ \hline \boldsymbol{x}_{2,1} \\ \hline \boldsymbol{x}_{2,2} \\ \hline \vdots \\ \hline \boldsymbol{x}_{2,m} \end{array}\right] \quad \text{and} \quad \boldsymbol{A}^{-1} = \left[\begin{array}{c|c|c|c|c} \boldsymbol{A}^{11} & \boldsymbol{A}^{12,1} & \boldsymbol{A}^{12,2} & \cdots & \boldsymbol{A}^{12,m} \\ \hline \boldsymbol{A}^{12,1\,T} & \boldsymbol{A}^{22,1} & \times & \cdots & \times \\ \hline \boldsymbol{A}^{12,2\,T} & \times & \boldsymbol{A}^{22,2} & \cdots & \times \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{A}^{12,m\,T} & \times & \times & \cdots & \boldsymbol{A}^{22,m} \end{array}\right] \tag{S.2}$$

with $\times$ denoting sub-blocks that are not of interest. The SolveTwoLevelSparseLeastSquares algorithm is given in Algorithm S.2.

S.5 Derivation of Result 1

The full conditional density function of $(\boldsymbol{\beta}, \boldsymbol{u})$ satisfies

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

Note that $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{u}', \sigma^2)$ can be expressed as the

$$N\left(\overset{\blacktriangle}{\boldsymbol{C}}\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}, \sigma^2\boldsymbol{I}\right)$$

density function in the vector $\overset{\blacktriangle}{\boldsymbol{r}}'$, where

$$\overset{\blacktriangle}{\boldsymbol{C}} \equiv \left[\underset{1\le i\le m}{\text{stack}}(\overset{\blacktriangle}{\boldsymbol{X}}_i) \quad \underset{1\le i\le m}{\text{blockdiag}}(\overset{\blacktriangle}{\boldsymbol{Z}}_i)\right] \quad \text{and} \quad \overset{\blacktriangle}{\boldsymbol{r}}' \equiv \underset{1\le i\le m}{\text{stack}}\left\{\overset{\blacktriangle}{\boldsymbol{y}}_i - \underset{1\le i'\le m'}{\text{stack}}(\boldsymbol{Z}'_{ii'}\boldsymbol{u}'_{i'})\right\}.$$

Also,

$$\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \quad \text{is the} \quad N\left(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma} \end{bmatrix}\right)$$

density function in the vector $(\boldsymbol{\beta}, \boldsymbol{u})$. Then, under product restriction II, standard quadratic form manipulations lead to the optimal $\mathfrak{q}$-density function of $(\boldsymbol{\beta}, \boldsymbol{u})$ being that of the $N(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})})$ distribution with updates

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow (\overset{\blacktriangle}{\boldsymbol{C}}{}^T\overset{\blacktriangle}{\boldsymbol{R}}{}^{-1}_{\text{MFVB}}\overset{\blacktriangle}{\boldsymbol{C}} + \overset{\blacktriangle}{\boldsymbol{D}}_{\text{MFVB}})^{-1}(\overset{\blacktriangle}{\boldsymbol{C}}{}^T\overset{\blacktriangle}{\boldsymbol{R}}{}^{-1}_{\text{MFVB}}\overset{\blacktriangle}{\boldsymbol{r}}{}'_{\text{MFVB}} + \overset{\blacktriangle}{\boldsymbol{o}}_{\text{MFVB}}) \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow (\overset{\blacktriangle}{\boldsymbol{C}}{}^T\overset{\blacktriangle}{\boldsymbol{R}}{}^{-1}_{\text{MFVB}}\overset{\blacktriangle}{\boldsymbol{C}} + \overset{\blacktriangle}{\boldsymbol{D}}_{\text{MFVB}})^{-1}.$$

Algorithm S.2 SOLVETWOLEVELSPARSELEASTSQUARES *for solving the two-level sparse matrix least squares problem: minimise $\|\boldsymbol{b}-\boldsymbol{B}\,\boldsymbol{x}\|^2$ in $\boldsymbol{x}$ and sub-blocks of $\boldsymbol{A}^{-1}$ corresponding to the non-zero sub-blocks of $\boldsymbol{A}=\boldsymbol{B}^T\boldsymbol{B}$. The sub-block notation is given by (S.1) and (S.2).*

Inputs: $\big\{\big(\boldsymbol{b}_i(\tilde{n}_i\times 1),\ \boldsymbol{B}_i(\tilde{n}_i\times p),\ \overset{\bullet}{\boldsymbol{B}}_i(\tilde{n}_i\times q)\big):\ 1\le i\le m\big\}$

$\boldsymbol{\omega}_3\longleftarrow$ NULL ; $\boldsymbol{\Omega}_4\longleftarrow$ NULL

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

Decompose $\overset{\bullet}{\boldsymbol{B}}_i=\boldsymbol{Q}_i\begin{bmatrix}\boldsymbol{R}_i\\ \boldsymbol{0}\end{bmatrix}$ such that $\boldsymbol{Q}_i^{-1}=\boldsymbol{Q}_i^T$ and $\boldsymbol{R}_i$ is upper-triangular.

$\boldsymbol{c}_{0i}\longleftarrow\boldsymbol{Q}_i^T\boldsymbol{b}_i$; $\boldsymbol{C}_{0i}\longleftarrow\boldsymbol{Q}_i^T\boldsymbol{B}_i$

$\boldsymbol{c}_{1i}\longleftarrow$ first q rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{c}_{2i}\longleftarrow$ remaining rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{\omega}_3\longleftarrow\begin{bmatrix}\boldsymbol{\omega}_3\\ \boldsymbol{c}_{2i}\end{bmatrix}$

$\boldsymbol{C}_{1i}\longleftarrow$ first q rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{C}_{2i}\longleftarrow$ remaining rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{\Omega}_4\longleftarrow\begin{bmatrix}\boldsymbol{\Omega}_4\\ \boldsymbol{C}_{2i}\end{bmatrix}$

Decompose $\boldsymbol{\Omega}_4=\boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{bmatrix}$ such that $\boldsymbol{Q}^{-1}=\boldsymbol{Q}^T$ and $\boldsymbol{R}$ is upper-triangular.

$\boldsymbol{c}\longleftarrow$ first p rows of $\boldsymbol{Q}^T\boldsymbol{\omega}_3$; $\boldsymbol{x}_1\longleftarrow\boldsymbol{R}^{-1}\boldsymbol{c}$; $\boldsymbol{A}^{11}\longleftarrow\boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$

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

$\boldsymbol{x}_{2,i}\longleftarrow\boldsymbol{R}_i^{-1}(\boldsymbol{c}_{1i}-\boldsymbol{C}_{1i}\boldsymbol{x}_1)$; $\boldsymbol{A}^{12,i}\longleftarrow-\boldsymbol{A}^{11}(\boldsymbol{R}_i^{-1}\boldsymbol{C}_{1i})^T$

$\boldsymbol{A}^{22,i}\longleftarrow\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i^{-T}-\boldsymbol{C}_{1i}\boldsymbol{A}^{12,i})$

Output: $\Big(\boldsymbol{x}_1,\boldsymbol{A}^{11},\big\{\big(\boldsymbol{x}_{2,i},\boldsymbol{A}^{22,i},\boldsymbol{A}^{12,i}\big):\ 1\le i\le m\big\}\Big)$

Here $\overset{\blacktriangle}{\boldsymbol{R}}_{\text{MFVB}}\equiv\mu_{\mathsf{q}(1/\sigma^2)}^{-1}\boldsymbol{I}$,

$$\overset{\blacktriangle}{\boldsymbol{D}}_{\text{MFVB}}\equiv\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}_m\otimes\boldsymbol{M}_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})}\end{bmatrix},\quad \overset{\blacktriangle}{\boldsymbol{o}}_{\text{MFVB}}\equiv\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\\ \boldsymbol{0}\end{bmatrix}$$

and

$$\overset{\blacktriangle}{\boldsymbol{r}}'_{\text{MFVB}}\equiv\underset{1\le i\le m}{\text{stack}}\Big\{\overset{\blacktriangle}{\boldsymbol{y}}_i-\underset{1\le i'\le m'}{\text{stack}}\big(\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\big)\Big\}.$$

If $\boldsymbol{b}$ and $\boldsymbol{B}$ are defined according to (S.1) and the matrices $\boldsymbol{b}_i$, $\boldsymbol{B}_i$ and $\overset{\bullet}{\boldsymbol{B}}_i$ are defined as in Result 1 then

$$\boldsymbol{B}^T\boldsymbol{b}=\overset{\blacktriangle}{\boldsymbol{C}}^T\overset{\blacktriangle}{\boldsymbol{R}}_{\text{MFVB}}^{-1}\overset{\blacktriangle}{\boldsymbol{r}}'_{\text{MFVB}}+\overset{\blacktriangle}{\boldsymbol{o}}_{\text{MFVB}}\quad\text{and}\quad\boldsymbol{B}^T\boldsymbol{B}=\overset{\blacktriangle}{\boldsymbol{C}}^T\overset{\blacktriangle}{\boldsymbol{R}}_{\text{MFVB}}^{-1}\overset{\blacktriangle}{\boldsymbol{C}}+\overset{\blacktriangle}{\boldsymbol{D}}_{\text{MFVB}}.$$

Therefore, with this assignment of $\boldsymbol{b}_i$, $\boldsymbol{B}_i$ and $\overset{\bullet}{\boldsymbol{B}}_i$, the $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ update corresponds to the least squares solution $\boldsymbol{x}=(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{b}$ and the updates of the sub-blocks of $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ listed (12) correspond to the sub-blocks of $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$ in the positions where $\boldsymbol{B}^T\boldsymbol{B}$ has non-zero sub-blocks.

S.6 Derivation of Result 2

Result 2 uses the following re-ordering of the overall design matrix:

$$\widetilde{\boldsymbol{C}}\equiv\Big[\boldsymbol{X}\ \ \underset{1\le i\le m}{\text{stack}}(\overset{\blacksquare}{\boldsymbol{Z}}'_i)\ \ \underset{1\le i\le m}{\text{blockdiag}}(\overset{\blacktriangle}{\boldsymbol{Z}}_i)\Big].$$

rather than $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$ in the generalized ridge regression expressions of Section 3. This re-ordering involves the $\mathfrak{q}$-density parameters of $\boldsymbol{u}'$ preceding those of $\boldsymbol{u}$ and is brought about by our $m \geq m'$ convention throughout this article and the requirement that the potentially very large

$$\underset{1\leq i\leq m}{\text{blockdiag}}(\overset{\blacktriangle}{\boldsymbol{Z}}_i)$$

appears on the right for embedding within the two-level sparse least squares infrastructure of Nolan & Wand (2019) and Nolan *et al.* (2019). The re-ordering means that (9) gets replaced by

$$\begin{aligned}
&\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}',\boldsymbol{u})} \longleftarrow (\widetilde{\boldsymbol{C}}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\widetilde{\boldsymbol{C}} + \widetilde{\boldsymbol{D}}_{\text{MFVB}})^{-1}(\widetilde{\boldsymbol{C}}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{y} + \boldsymbol{o}_{\text{MFVB}}) \quad \text{and} \\
&\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}',\boldsymbol{u})} \longleftarrow (\widetilde{\boldsymbol{C}}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\widetilde{\boldsymbol{C}} + \widetilde{\boldsymbol{D}}_{\text{MFVB}})^{-1}
\end{aligned}$$

where

$$\widetilde{\boldsymbol{D}}_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{m'} \otimes \boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} \end{bmatrix}$$

has the $\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}$ matrices appearing before the $\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}$ matrices due to the switch in the ordering of the random effects vectors.

If $\boldsymbol{b}$ and $\boldsymbol{B}$ are defined according to (S.1) with the matrices $\boldsymbol{b}_i$, $\boldsymbol{B}_i$ and $\overset{\bullet}{\boldsymbol{B}}_i$ defined as in Result 2 then straightforward matrix algebra can be used to show that

$$\boldsymbol{B}^T\boldsymbol{b} = \widetilde{\boldsymbol{C}}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{y} + \boldsymbol{o}_{\text{MFVB}} \quad \text{and} \quad \boldsymbol{B}^T\boldsymbol{B} = \widetilde{\boldsymbol{C}}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\widetilde{\boldsymbol{C}} + \widetilde{\boldsymbol{D}}_{\text{MFVB}}.$$

Therefore, with this assignment of $\boldsymbol{b}_i$, $\boldsymbol{B}_i$ and $\overset{\bullet}{\boldsymbol{B}}_i$, the $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}',\boldsymbol{u})}$ update corresponds to the least squares solution $\boldsymbol{x} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{b}$ and the updates of the sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}',\boldsymbol{u})}$ listed (11) correspond to the sub-blocks of $(\boldsymbol{B}^T\boldsymbol{B})^{-1}$ in the positions where $\boldsymbol{B}^T\boldsymbol{B}$ has non-zero sub-blocks.

S.7 Marginal Log-Likelihood Lower Bound and Derivation

The logarithmic form of the variational lower bound on the marginal log-likelihood, corresponding to model (1) with prior specification (B) and product restriction III is

$$\begin{aligned}
\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q}) &= E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}},\boldsymbol{A}_{\boldsymbol{\Sigma}'},\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'\right)\right. \\
&\qquad \left. - \log \mathfrak{q}^*\left(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}},\boldsymbol{A}_{\boldsymbol{\Sigma}'},\sigma^2,\boldsymbol{\Sigma},\boldsymbol{\Sigma}'\right)\right\} \\
&= E_{\mathfrak{q}}\left\{\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2\right)\right\} + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}'|\boldsymbol{\Sigma},\boldsymbol{\Sigma}'\right)\right\} \\
&\quad - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}'\right)\right\} + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\sigma^2|a_{\sigma^2}\right)\right\} - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(\sigma^2\right)\right\} \\
&\quad + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(a_{\sigma^2}\right)\right\} - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(a_{\sigma^2}\right)\right\} + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}}\right)\right\} \\
&\quad - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(\boldsymbol{\Sigma}\right)\right\} + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\boldsymbol{A}_{\boldsymbol{\Sigma}}\right)\right\} - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(\boldsymbol{A}_{\boldsymbol{\Sigma}}\right)\right\} \\
&\quad + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\boldsymbol{\Sigma}'|\boldsymbol{A}_{\boldsymbol{\Sigma}'}\right)\right\} - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(\boldsymbol{\Sigma}'\right)\right\} + E_{\mathfrak{q}}\left\{\log \mathfrak{p}\left(\boldsymbol{A}_{\boldsymbol{\Sigma}'}\right)\right\} \\
&\quad - E_{\mathfrak{q}}\left\{\log \mathfrak{q}^*\left(\boldsymbol{A}_{\boldsymbol{\Sigma}'}\right)\right\}.
\end{aligned}$$

The first of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is

$$
\begin{aligned}
E_{\mathfrak{q}}\left\{\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2\right)\right\} &= -\tfrac{1}{2}n_{\bullet\bullet}\log(2\pi)-\tfrac{1}{2}n_{\bullet\bullet}E_{\mathfrak{q}}\left\{\log(\sigma^2)\right\}\\
&\quad -\tfrac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2)}\sum_{i=1}^{m}\sum_{i'=1}^{m'}\left\{\left|\left|\boldsymbol{y}_{ii'}-\boldsymbol{X}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}-\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\right|\right|^2\right.\\
&\qquad +\text{tr}\left(\boldsymbol{X}_{ii'}^T\boldsymbol{X}_{ii'}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right)+\text{tr}\left(\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}_{ii'}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)\\
&\qquad +\text{tr}\left(\left(\boldsymbol{Z}'_{ii'}\right)^T\boldsymbol{Z}'_{ii'}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\right)\\
&\qquad +2\text{tr}\left[\boldsymbol{Z}_{ii'}^T\boldsymbol{X}_{ii'}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}\right]\\
&\qquad +2\text{tr}\left[\left(\boldsymbol{Z}'_{ii'}\right)^T\boldsymbol{X}_{ii'}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}\right]\\
&\qquad \left.+2\text{tr}\left[\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}'_{ii'}E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}\right]\right\}.
\end{aligned}
$$

Under product restrictions I and II, $E_{\mathfrak{q}}\left\{\mathfrak{p}\left(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}',\sigma^2\right)\right\}$ simplify further as we have

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}=\boldsymbol{O},\quad 1\le i'\le m',$$

and

$$E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}=\boldsymbol{O},\quad 1\le i\le m, 1\le i'\le m'.$$

Under product restriction I we also have

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}=\boldsymbol{O},\quad 1\le i\le m.$$

The second of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is

$$E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}'\mid\boldsymbol{\Sigma},\boldsymbol{\Sigma}')\}]=E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{\beta})\}+\log\{\mathfrak{p}(\boldsymbol{u}|\boldsymbol{\Sigma})\}+\log\{\mathfrak{p}(\boldsymbol{u}'|\boldsymbol{\Sigma}')\}]$$

$$
\begin{aligned}
&= -\tfrac{1}{2}(p+mq+m'q')\log(2\pi)-\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|-\frac{m}{2}E_{\mathfrak{q}}\left\{\log|\boldsymbol{\Sigma}|\right\}-\frac{m'}{2}E_{\mathfrak{q}}\left\{\log|\boldsymbol{\Sigma}'|\right\}\\
&\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right\}\right)\\
&\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)\right\}\right)\\
&\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathfrak{q}\left((\boldsymbol{\Sigma}')^{-1}\right)}\left\{\sum_{i'=1}^{m'}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\right)\right\}\right).
\end{aligned}
$$

The third of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is the negative of

$$E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')\}] = -\tfrac{1}{2}(p+mq+m'q')-\tfrac{1}{2}(p+mq+m'q')\log(2\pi)-\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')}|.$$

The fourth of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(\sigma^2|a_{\sigma^2})\}] &= E_{\mathfrak{q}}\left(\log\left[\frac{\{1/(2a_{\sigma^2})\}^{\nu_{\sigma^2}/2}}{\Gamma(\nu_{\sigma^2}/2)}(\sigma^2)^{-(\nu_{\sigma^2}/2)-1}\exp\{-1/(2a_{\sigma^2}\sigma^2)\}\right]\right)\\
&= -\tfrac{1}{2}\nu_{\sigma^2}\,E_{\mathfrak{q}}\{\log(2a_{\sigma^2})\}-\log\{\Gamma(\tfrac{1}{2}\nu_{\sigma^2})\}-(\tfrac{1}{2}\nu_{\sigma^2}+1)E_{\mathfrak{q}}\{\log(\sigma^2)\}\\
&\quad -\tfrac{1}{2}\mu_{\mathfrak{q}(1/a_{\sigma^2})}\mu_{\mathfrak{q}(1/\sigma^2)}.
\end{aligned}
$$

The fifth of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is the negative of

$$\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(\sigma^2)\}] &= E_{\mathfrak{q}}\left(\log\left[\frac{\{\lambda_{\mathfrak{q}(\sigma^2)}/2\}^{\xi_{\mathfrak{q}(\sigma^2)}/2}}{\Gamma(\xi_{\mathfrak{q}(\sigma^2)}/2)}(\sigma^2)^{-(\xi_{\mathfrak{q}(\sigma^2)}/2)-1}\exp\{-\lambda_{\mathfrak{q}(\sigma^2)}/(2\sigma^2)\}\right]\right)\\
&= \tfrac{1}{2}\xi_{\mathfrak{q}(\sigma^2)}\,\log(\lambda_{\mathfrak{q}(\sigma^2)}/2)-\log\{\Gamma(\tfrac{1}{2}\xi_{\mathfrak{q}(\sigma^2)})\}-(\tfrac{1}{2}\xi_{\mathfrak{q}(\sigma^2)}+1)E_{\mathfrak{q}}\{\log(\sigma^2)\}\\
&\quad -\tfrac{1}{2}\lambda_{\mathfrak{q}(\sigma^2)}\mu_{\mathfrak{q}(1/\sigma^2)}.
\end{aligned}$$

The sixth of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is

$$\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(a_{\sigma^2})\}] &= E_{\mathfrak{q}}\left(\log\left[\frac{\{1/(2\nu_{\sigma^2}s_{\sigma^2}^2)\}^{1/2}}{\Gamma(1/2)}a_{\sigma^2}^{-(1/2)-1}\exp\{-1/(2\nu_{\sigma^2}s_{\sigma^2}^2a_{\sigma^2})\}\right]\right)\\
&= -\tfrac{1}{2}\,\log(2\nu_{\sigma^2}s_{\sigma^2}^2)-\log\{\Gamma(\tfrac{1}{2})\}-(\tfrac{1}{2}+1)E_{\mathfrak{q}}\{\log(a_{\sigma^2})\}-\{1/(2\nu_{\sigma^2}s_{\sigma^2}^2)\}\mu_{\mathfrak{q}(1/a_{\sigma^2})}.
\end{aligned}$$

The seventh of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is the negative of

$$\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(a_{\sigma^2})\}] &= E_{\mathfrak{q}}\left(\log\left[\frac{\{\lambda_{\mathfrak{q}(a_{\sigma^2})}/2\}^{\xi_{\mathfrak{q}(a_{\sigma^2})}/2}}{\Gamma(\xi_{\mathfrak{q}(a_{\sigma^2})}/2)}(a_{\sigma^2})^{-(\xi_{\mathfrak{q}(a_{\sigma^2})}/2)-1}\exp\{-\lambda_{\mathfrak{q}(a_{\sigma^2})}/(2a_{\sigma^2})\}\right]\right)\\
&= \tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})}\,\log(\lambda_{\mathfrak{q}(a_{\sigma^2})}/2)-\log\{\Gamma(\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})})\}-(\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})}+1)E_{\mathfrak{q}}\{\log(a_{\sigma^2})\}\\
&\quad -\tfrac{1}{2}\lambda_{\mathfrak{q}(a_{\sigma^2})}\mu_{\mathfrak{q}(1/a_{\sigma^2})}.
\end{aligned}$$

The eighth of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is

$$\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}})\}] &= E_{\mathfrak{q}}\left(\frac{|\boldsymbol{A}_{\boldsymbol{\Sigma}}|^{-\frac{1}{2}(\nu_{\boldsymbol{\Sigma}}+q-1)}|\boldsymbol{\Sigma}|^{-\frac{1}{2}(\nu_{\boldsymbol{\Sigma}}+2q)}}{2^{\frac{q}{2}(\nu_{\boldsymbol{\Sigma}}+2q-1)}\pi^{\frac{q}{4}(q-1)}\prod_{j=1}^{q}\Gamma(\frac{1}{2}(\nu_{\boldsymbol{\Sigma}}+2q-j))}\exp\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}^{-1})\}\right)\\
&= -\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+q-1)E_{\mathfrak{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}}|\}-\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+2q)E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}|\}\\
&\quad -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})})-\frac{q}{2}(\nu_{\boldsymbol{\Sigma}}+2q-1)\log(2)-\frac{q}{4}(q-1)\log(\pi)\\
&\quad -\sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+2q-j)).
\end{aligned}$$

The ninth of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is the negative of

$$\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{\Sigma})\}] &= E_{\mathfrak{q}}\left(\frac{|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}|^{\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}-q+1)}|\boldsymbol{\Sigma}|^{-\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+2)}}{2^{\frac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+1)}\pi^{\frac{q}{4}(q-1)}\prod_{j=1}^{q}\Gamma(\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+2-j))}\exp\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}^{-1})\}\right)\\
&= \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}-q+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}|-\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+2)E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}|\}-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})})\\
&\quad -\frac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+1)\log(2)-\frac{q}{4}(q-1)\log(\pi)-\sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+2-j)).
\end{aligned}$$

The tenth of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is

$$\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}})\}] &= E_{\mathfrak{q}}\left(\frac{|\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}|^{\frac{1}{2}(2-q)}|\boldsymbol{A}_{\boldsymbol{\Sigma}}|^{-3/2}}{2^{q}\pi^{\frac{q}{4}(q-1)}\prod_{j=1}^{q}\Gamma(\frac{1}{2}(3-j))}\exp\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})\}\right)\\
&= -\tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}})-\tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s_{\boldsymbol{\Sigma},j}^2)-\frac{3}{2}E_{\mathfrak{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}}|\}\\
&\quad -\tfrac{1}{2}\sum_{j=1}^{q}1/(\nu_{\boldsymbol{\Sigma}}s_{\boldsymbol{\Sigma},j}^2)\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\right)_{jj}-q\log(2)-\frac{q}{4}(q-1)\log(\pi)\\
&\quad -\sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(3-j)).
\end{aligned}$$

The eleventh of the $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ terms is the negative of

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{A_\Sigma})\}] &= E_{\mathfrak{q}}\left(\frac{|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A_\Sigma})}|^{\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}-q+1)}|\boldsymbol{A_\Sigma}|^{-\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}+2)}}{2^{\frac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}+1)}\pi^{\frac{q}{4}(q-1)}\prod_{j=1}^{q}\Gamma(\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}+2-j))}\exp\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A_\Sigma})}^{-1}\boldsymbol{A_\Sigma}^{-1})\}\right)\\
&= \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}-q+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A_\Sigma})}| - \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}+2)E_{\mathfrak{q}}\{\log|\boldsymbol{A_\Sigma}|\} - \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A_\Sigma})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A_\Sigma}^{-1})})\\
&\quad -\frac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}+1)\log(2) - \frac{q}{4}(q-1)\log(\pi) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_\Sigma})}+2-j)).
\end{aligned}
$$

The remaining four terms of $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ are

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{\Sigma}'|\boldsymbol{A_{\Sigma'}})\}] &= -\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+q'-1)E_{\mathfrak{q}}\{\log|\boldsymbol{A_{\Sigma'}}|\} - \tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+2q')E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}'|\}\\
&\quad -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A_{\Sigma'}})^{-1})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma})^{-1})}) - \frac{q'}{2}(\nu_{\boldsymbol{\Sigma}'}+2q'-1)\log(2) - \frac{q'}{4}(q'-1)\log(\pi)\\
&\quad -\sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+2q'-j)),
\end{aligned}
$$

the negative of

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{\Sigma}')\}] &= \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}-q'+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}| - \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}+2)E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}'|\} - \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})})\\
&\quad -\frac{q'}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}+1)\log(2) - \frac{q'}{4}(q'-1)\log(\pi) - \sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}+2-j)),
\end{aligned}
$$

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{A_{\Sigma'}})\}] &= -\tfrac{1}{2}q'(2-q')\log(\nu_{\boldsymbol{\Sigma}'}) - \tfrac{1}{2}(2-q')\sum_{j'=1}^{q'}\log(s_{\boldsymbol{\Sigma},j}^2) - \frac{3}{2}E_{\mathfrak{q}}\{\log|\boldsymbol{A_\Sigma}|\}\\
&\quad -\tfrac{1}{2}\sum_{j=1}^{q'}1/(\nu_{\boldsymbol{\Sigma}'}s_{\boldsymbol{\Sigma},j}^2)\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A_{\Sigma'}})^{-1})}\right)_{jj} - q'\log(2) - \frac{q'}{4}(q'-1)\log(\pi)\\
&\quad -\sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(3-j)),
\end{aligned}
$$

and the negative of

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})\}] &= \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})}-q'+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})}| - \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})}+2)E_{\mathfrak{q}}\{\log|\boldsymbol{A_{\Sigma'}}|\}\\
&\quad -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A_{\Sigma'}})^{-1})}) - \frac{q'}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})}+1)\log(2) - \frac{q'}{4}(q'-1)\log(\pi)\\
&\quad -\sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A_{\Sigma'}})}+2-j)).
\end{aligned}
$$

In the summation of each of these $\log \underline{\mathfrak{p}}(\boldsymbol{x};q)$ terms, note that the coefficient of $E_{\mathfrak{q}}\{\log(\sigma^2)\}$ is

$$-\tfrac{1}{2}\,n_{\bullet\bullet} - \tfrac{1}{2}\nu_{\sigma^2} - 1 + \tfrac{1}{2}\xi_{\mathfrak{q}(\sigma^2)} + 1 = -\tfrac{1}{2}\,n_{\bullet\bullet} - \tfrac{1}{2}\nu_{\sigma^2} - 1 + \tfrac{1}{2}(\nu_{\sigma^2}+n_{\bullet\bullet}) + 1 = 0.$$

The coefficient of $E_{\mathfrak{q}}\{\log(a_{\sigma^2})\}$ is

$$-\tfrac{1}{2}\nu_{\sigma^2} - (\tfrac{1}{2}+1) + \tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})} + 1 = -\tfrac{1}{2}\nu_{\sigma^2} - (\tfrac{1}{2}+1) + \tfrac{1}{2}(\nu_{\sigma^2}+1) + 1 = 0.$$

The coefficient of $E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}|\}$ is

$$-\frac{m}{2} - \tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+2q) + \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+2) = -\tfrac{1}{2}(m+\nu_{\boldsymbol{\Sigma}}+2q) + \tfrac{1}{2}(m+\nu_{\boldsymbol{\Sigma}}+2q) = 0.$$

The coefficient of $E_{\mathsf{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}}|\}$ is

$$-\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+q-1)-\frac{3}{2}+\tfrac{1}{2}(\xi_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}+2)=-\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+q+2)+\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+q+2)=0.$$

The coefficient of $E_{\mathsf{q}}\{\log|\boldsymbol{\Sigma}'|\}$ is

$$-\frac{m'}{2}-\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+2q')+\tfrac{1}{2}(\xi_{\mathsf{q}(\boldsymbol{\Sigma}')}+2)=-\tfrac{1}{2}(m'+\nu_{\boldsymbol{\Sigma}'}+2q')+\tfrac{1}{2}(m'+\nu_{\boldsymbol{\Sigma}'}+2q')=0.$$

The coefficient of $E_{\mathsf{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}'}|\}$ is

$$-\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+q'-1)-\frac{3}{2}+\tfrac{1}{2}(\xi_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})}+2)=-\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+q'+2)+\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+q'+2)=0.$$

Therefore, the terms in $E_{\mathsf{q}}\{\log(\sigma^2)\}$, $E_{\mathsf{q}}\{\log(a)\}$, $E_{\mathsf{q}}\{\log|\boldsymbol{\Sigma}|\}$ and $E_{\mathsf{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}}|\}$ can be dropped and we then have

$$\log\underline{\mathfrak{p}}(\boldsymbol{y};\mathsf{q})=\sum_{i=1}^{15}T_i$$

where

$$\begin{aligned}
T_1 &= -\tfrac{1}{2}n_{\bullet\bullet}\log(2\pi)\\
&\quad -\tfrac{1}{2}\mu_{\mathsf{q}(1/\sigma^2)}\sum_{i=1}^{m}\sum_{i'=1}^{m'}\Bigg\{\Big|\Big|\boldsymbol{y}_{ii'}-\boldsymbol{X}_{ii'}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}-\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\Big|\Big|^2\\
&\quad +\text{tr}\left(\boldsymbol{X}_{ii'}^T\boldsymbol{X}_{ii'}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}\right)+\text{tr}\left(\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}_{ii'}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}\right)+\text{tr}\left(\left(\boldsymbol{Z}'_{ii'}\right)^T\boldsymbol{Z}'_{ii'}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\right)\\
&\quad +2\text{tr}\left[\boldsymbol{Z}_{ii'}^T\boldsymbol{X}_{ii'}E_{\mathsf{q}}\left\{\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}\right)\left(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}\right)^T\right\}\right]\\
&\quad +2\text{tr}\left[\left(\boldsymbol{Z}'_{ii'}\right)^T\boldsymbol{X}_{ii'}E_{\mathsf{q}}\left\{\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}\right)\left(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\right)^T\right\}\right]\\
&\quad +2\text{tr}\left[\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}'_{ii'}E_{\mathsf{q}}\left\{\left(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}\right)\left(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\right)^T\right\}\right]\Bigg\},\\
T_2 &= -\tfrac{1}{2}(p+mq+m'q')\log(2\pi)-\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|\\
&\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^T+\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}\right\}\right)\\
&\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}^T+\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}\right)\right\}\right)\\
&\quad -\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}')^{-1})}\left\{\sum_{i'=1}^{m'}\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}'_{i'})}^T+\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}'_{i'})}\right)\right\}\right),\\
T_3 &= \tfrac{1}{2}(p+mq+m'q')+\tfrac{1}{2}(p+mq+m'q')\log(2\pi)+\tfrac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')}\right|,\\
T_4 &= \tfrac{1}{2}\nu_{\sigma^2}\log(2)-\log\{\Gamma(\tfrac{1}{2}\nu_{\sigma^2})\}-\tfrac{1}{2}\mu_{\mathsf{q}(1/a_{\sigma^2})}\mu_{\mathsf{q}(1/\sigma^2)},\\
T_5 &= -\tfrac{1}{2}\xi_{\mathsf{q}(\sigma^2)}\ \log(\lambda_{\mathsf{q}(\sigma^2)}/2)+\log\{\Gamma(\tfrac{1}{2}\xi_{\mathsf{q}(\sigma^2)})\}+\tfrac{1}{2}\lambda_{\mathsf{q}(\sigma^2)}\mu_{\mathsf{q}(1/\sigma^2)},\\
T_6 &= -\tfrac{1}{2}\ \log(2\nu_{\sigma^2}s_{\sigma^2}^2)-\log\{\Gamma(\tfrac{1}{2})\}-\{1/(2\nu_{\sigma^2}s_{\sigma^2}^2)\}\mu_{\mathsf{q}(1/a_{\sigma^2})}
\end{aligned}$$

$$T_7 = -\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})}\,\log(\lambda_{\mathfrak{q}(a_{\sigma^2})}/2) + \log\{\Gamma(\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})})\} + \tfrac{1}{2}\lambda_{\mathfrak{q}(a_{\sigma^2})}\mu_{\mathfrak{q}(1/a_{\sigma^2})},$$

$$T_8 = -\tfrac{1}{2}\text{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}) - \frac{q}{2}(\nu_{\boldsymbol{\Sigma}} + 2q - 1)\log(2) - \frac{q}{4}(q-1)\log(\pi) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}} + 2q - j)),$$

$$T_9 = -\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} - q + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}| + \tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}) + \frac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} + 1)\log(2), + \frac{q}{4}(q-1)\log(\pi) + \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} + 2 - j)),$$

$$T_{10} = -\tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) - \tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s^2_{\boldsymbol{\Sigma},j}) - \tfrac{1}{2}\sum_{j=1}^{q}1/(\nu_{\boldsymbol{\Sigma}}s^2_{\boldsymbol{\Sigma},j})\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\right)_{jj} - q\log(2) - \frac{q}{4}(q-1)\log(\pi) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(3-j)),$$

$$T_{11} = -\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} - q + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}| + \tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}) + \frac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} + 1)\log(2) + \frac{q}{4}(q-1)\log(\pi) + \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} + 2 - j)),$$

$$T_{12} = -\tfrac{1}{2}\text{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}) - \frac{q'}{2}(\nu_{\boldsymbol{\Sigma}'} + 2q' - 1)\log(2) - \frac{q'}{4}(q'-1)\log(\pi) - \sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'} + 2q' - j)),$$

$$T_{13} = -\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')} - q' + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}| + \tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}) + \frac{q'}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')} + 1)\log(2), + \frac{q'}{4}(q'-1)\log(\pi) + \sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')} + 2 - j)),$$

$$T_{14} = -\tfrac{1}{2}q'(2-q')\log(\nu_{\boldsymbol{\Sigma}'}) - \tfrac{1}{2}(2-q')\sum_{j=1}^{q'}\log(s^2_{\boldsymbol{\Sigma},j}) - \tfrac{1}{2}\sum_{j=1}^{q'}1/(\nu_{\boldsymbol{\Sigma}'}s^2_{\boldsymbol{\Sigma},j})\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A}_{\boldsymbol{\Sigma}'})^{-1})}\right)_{jj} - q'\log(2) - \frac{q'}{4}(q'-1)\log(\pi) - \sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(3-j))$$

and
$$T_{15} = -\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})} - q' + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})}| + \tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A}_{\boldsymbol{\Sigma}'})^{-1})}) + \frac{q'}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})} + 1)\log(2) + \frac{q'}{4}(q'-1)\log(\pi) + \sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})} + 2 - j)).$$

Note that the component of $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ which does not get updated during the coordinate ascent iterations, except for the irreducible $\log\Gamma$ terms, and which we will call 'const' is:

$$
\begin{aligned}
\text{const} \;\equiv\; & -\tfrac{1}{2}n_{\bullet\bullet}\log(2\pi) - \tfrac{1}{2}(p+mq+m'q')\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \tfrac{1}{2}(p+mq+m'q') \\
& +\tfrac{1}{2}(p+mq+m'q')\log(2\pi) - \tfrac{1}{2}\nu_{\sigma^2}\log(2) + \tfrac{1}{2}(\xi_{\mathfrak{q}(\sigma^2)})\log(2) - \tfrac{1}{2}\log(2\nu_{\sigma^2}s^2_{\sigma^2}) \\
& -\tfrac{1}{2}q(\nu_{\boldsymbol{\Sigma}}+2q-1)\log(2) - \frac{q}{2}(q-1)\log(\pi) + \tfrac{1}{2}q(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+1)\log(2) + \frac{q}{2}(q-1)\log(\pi) \\
& -\tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) - \tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s^2_{\boldsymbol{\Sigma},j}) - q\log(2) + \tfrac{1}{2}q(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}+1)\log(2) \\
& -\tfrac{1}{2}q'(\nu_{\boldsymbol{\Sigma}'}+2q'-1)\log(2) - \frac{q'}{2}(q'-1)\log(\pi) + \tfrac{1}{2}q'(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}+1)\log(2) \\
& +\frac{q'}{2}(q'-1)\log(\pi) - \tfrac{1}{2}q'(2-q')\log(\nu_{\boldsymbol{\Sigma}'}) - \tfrac{1}{2}(2-q')\sum_{j=1}^{q'}\log(s^2_{\boldsymbol{\Sigma},j}) - q'\log(2) \\
& +\tfrac{1}{2}q'(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'})}+1)\log(2) - \log\Gamma(\tfrac{1}{2}) \\
=\; & -\tfrac{1}{2}\,(n_{\bullet\bullet}+1)\log(\pi) - \tfrac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| + \tfrac{1}{2}(p+mq+m'q') - \tfrac{1}{2}\log(\nu_{\sigma^2}) - \tfrac{1}{2}\log(s^2_{\sigma^2}) \\
& +\tfrac{1}{2}\,\{q(\nu_{\boldsymbol{\Sigma}}+q+m-1) + q'(\nu_{\boldsymbol{\Sigma}'}+q'+m'-1) - 1\}\log(2) \\
& -\tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) - \tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s^2_{\sigma^2}) - \tfrac{1}{2}q'(2-q')\log(\nu_{\boldsymbol{\Sigma}'}) - \tfrac{1}{2}(2-q')\sum_{j=1}^{q'}\log(s^2_{\sigma^2})
\end{aligned}
$$

Our final $\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ expression is then

$$
\begin{aligned}
\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q}) =\; & -\tfrac{1}{2}\,(n_{\bullet\bullet}+1)\log(\pi) - \tfrac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| + \tfrac{1}{2}(p+mq+m'q') - \tfrac{1}{2}\log(\nu_{\sigma^2}) - \tfrac{1}{2}\log(s^2_{\sigma^2}) \\
& +\tfrac{1}{2}\,\{q(\nu_{\boldsymbol{\Sigma}}+q+m-1) + q'(\nu_{\boldsymbol{\Sigma}'}+q'+m'-1) - 1\}\log(2) - \tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) \\
& -\tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s^2_{\sigma^2}) - \tfrac{1}{2}q'(2-q')\log(\nu_{\boldsymbol{\Sigma}'}) - \tfrac{1}{2}(2-q')\sum_{j=1}^{q'}\log(s^2_{\sigma^2}) - \log\{\Gamma(\tfrac{1}{2}\nu_{\sigma^2})\} \\
& -\tfrac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right\}\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')}| \\
& -\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)\right\}\right) - \tfrac{1}{2}\mu_{\mathfrak{q}(1/a_{\sigma^2})}\mu_{\mathfrak{q}(1/\sigma^2)} \\
& -\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}\left\{\sum_{i'=1}^{m'}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\right)\right\}\right) - \{1/(2\nu_{\sigma^2}s^2_{\sigma^2})\}\mu_{\mathfrak{q}(1/a_{\sigma^2})} \\
& -\tfrac{1}{2}\xi_{\mathfrak{q}(\sigma^2)}\;\log(\lambda_{\mathfrak{q}(\sigma^2)}/2) + \log\{\Gamma(\tfrac{1}{2}\xi_{\mathfrak{q}(\sigma^2)})\} + \tfrac{1}{2}\lambda_{\mathfrak{q}(\sigma^2)}\mu_{\mathfrak{q}(1/\sigma^2)} \\
& -\tfrac{1}{2}\text{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}}+2q-j)) - \tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}) \\
& -\tfrac{1}{2}\text{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}'}^{-1})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}) - \sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}'}+2q'-j)) - \tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}')^{-1})}) \\
& -\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})}\;\log(\lambda_{\mathfrak{q}(a_{\sigma^2})}/2) + \log\{\Gamma(\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})})\} + \tfrac{1}{2}\lambda_{\mathfrak{q}(a_{\sigma^2})}\mu_{\mathfrak{q}(1/a_{\sigma^2})} \\
& -\sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}+2-j)) + \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})}-q+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}|
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_{j=1}^{q'}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}+2-j))+\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}')}-q'+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}')}|\\
&-\tfrac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2)}\sum_{i=1}^{m}\sum_{i'=1}^{m'}\left\{\left|\left|\boldsymbol{y}_{ii'}-\boldsymbol{X}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}-\boldsymbol{Z}'_{ii'}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})}\right|\right|^2\right.\\
&\quad+\text{tr}\left(\boldsymbol{X}_{ii'}^T\boldsymbol{X}_{ii'}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right)+\text{tr}\left(\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}_{ii'}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)+\text{tr}((\boldsymbol{Z}'_{ii'})^T\boldsymbol{Z}'_{ii'}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})\\
&\quad+2\text{tr}\left[\boldsymbol{Z}_{ii'}^T\boldsymbol{X}_{ii'}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}\right]\\
&\quad+2\text{tr}\left[\left(\boldsymbol{Z}'_{ii'}\right)^T\boldsymbol{X}_{ii'}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}\right]\\
&\quad\left.+2\text{tr}\left[\boldsymbol{Z}_{ii'}^T\boldsymbol{Z}'_{ii'}E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}\right]\right\}.
\end{aligned}
$$

The $\log\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})$ expression simplifies under product restrictions I and II, since we have

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}=\boldsymbol{O},\quad 1\le i'\le m',$$

and

$$E_{\mathfrak{q}}\{(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})(\boldsymbol{u}'_{i'}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}'_{i'})})^T\}=\boldsymbol{O},\quad 1\le i\le m, 1\le i'\le m'.$$

Under product restriction I we also have

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}=\boldsymbol{O},\quad 1\le i\le m.$$

From Theorem 1 of Nolan & Wand (2018), the $\log|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')}|$ term has the following streamlined form:

$$
\begin{aligned}
\log|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')}|=&\log\left|\boldsymbol{A}^{11}\text{ component of }\mathcal{S}\text{ from Algorithm 2}\right|\\
&-\sum_{i=1}^{m}\log\left|\mu_{\mathfrak{q}(1/\sigma^2)}\overset{\blacktriangle}{\boldsymbol{Z}}{}_i^T\overset{\blacktriangle}{\boldsymbol{Z}}_i+\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\right|,
\end{aligned}
$$

under product restriction I, and

$$\log|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{u}')}|=\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right|+\sum_{i=1}^{m}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right|-\sum_{i=1}^{m}\log\left|\mu_{\mathfrak{q}(1/\sigma^2)}\overset{\blacktriangle}{\boldsymbol{Z}}{}_i^T\overset{\blacktriangle}{\boldsymbol{Z}}_i+\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\right|$$

under product restrictions II and III.

S.8 Streamlined Computing for Frequentist Inference

As an aside we point out that the approach used by Algorithm 3 for product restriction III, in which the $\mathfrak{q}$-density updates for the $(\boldsymbol{\beta},\boldsymbol{u}_{\text{all}})$ parameters are embedded within the SolveTwoLevelSparseLeastSquares infrastructure, can also be used for streamlined *frequentist* inference when m' is moderate in size. To the best of our knowledge, the results given here for efficient computation of the important sub-blocks of the relevant covariance matrix are novel.

The frequentist Gaussian response two-level linear mixed model with crossed random effects is

$$
\begin{aligned}
&\boldsymbol{y}_{ii'}|\boldsymbol{\beta},\boldsymbol{u}_i,\boldsymbol{u}'_{i'}\overset{\text{ind.}}{\sim}N(\boldsymbol{X}_{ii'}\boldsymbol{\beta}+\boldsymbol{Z}_{ii'}\,\boldsymbol{u}_i+\boldsymbol{Z}'_{ii'}\,\boldsymbol{u}'_{i'},\sigma^2\,\boldsymbol{I}),\\
&\boldsymbol{u}_i\overset{\text{ind.}}{\sim}N(\boldsymbol{0},\boldsymbol{\Sigma}),\ \boldsymbol{u}'_{i'}\overset{\text{ind.}}{\sim}N(\boldsymbol{0},\boldsymbol{\Sigma}'),\ 1\le i\le m,\ 1\le i\le m'.
\end{aligned}
\qquad \text{(S.3)}
$$

The best linear unbiased predictor of $[\boldsymbol{\beta}^T\ \boldsymbol{u}^T]^T$ and its corresponding covariance matrix are

$$\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{u}}\end{bmatrix} = (\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1}\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{y}$$
$$\text{and}\quad \text{Cov}\left(\begin{bmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\boldsymbol{u}}-\boldsymbol{u}\end{bmatrix}\right) = (\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1} \tag{S.4}$$

where $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$, with $\boldsymbol{X}$ and $\boldsymbol{Z}$ as defined in Section 2.1.

$$\boldsymbol{D}_{\text{BLUP}} \equiv \begin{bmatrix}\boldsymbol{O} & \boldsymbol{O}\\ \boldsymbol{O} & \begin{bmatrix}\boldsymbol{I}_m\otimes\boldsymbol{\Sigma}^{-1} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}_{m'}\otimes(\boldsymbol{\Sigma}')^{-1}\end{bmatrix}\end{bmatrix} \quad\text{and}\quad \boldsymbol{R}_{\text{BLUP}} \equiv \sigma^2\boldsymbol{I}.$$

Note that the following sub-blocks are required for adding pointwise confidence intervals to mean estimates:

$$\begin{gathered}\text{Cov}(\widehat{\boldsymbol{\beta}}),\quad \text{Cov}(\widehat{\boldsymbol{u}}_i-\boldsymbol{u}_i),\quad \text{Cov}(\widehat{\boldsymbol{u}}'_{i'}-\boldsymbol{u}'_{i'}),\\ E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i-\boldsymbol{u}_i)^T\},\quad E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}'_{i'}-\boldsymbol{u}'_{i'})^T\}\quad\text{and}\quad E\{(\widehat{\boldsymbol{u}}_i-\boldsymbol{u}_i)(\widehat{\boldsymbol{u}}'_{i'}-\boldsymbol{u}'_{i'})^T\}\end{gathered} \tag{S.5}$$

for $1\le i\le m$ and $1\le i'\le m'$.

Result S.1. *Computation of $[\widehat{\boldsymbol{\beta}}^T\ \ \widehat{\boldsymbol{u}}^T]^T$ and each of the sub-blocks of Cov$([\widehat{\boldsymbol{\beta}}\ \ \widehat{\boldsymbol{u}}-\boldsymbol{u}]^T)$ listed in (S.5) are expressible as the two-level sparse matrix least squares form:*

$$\left\|\boldsymbol{b}-\boldsymbol{B}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right\|^2$$

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

$$\boldsymbol{b}_i \equiv \begin{bmatrix}\sigma^{-1}\boldsymbol{y}_i\\ \boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\quad \boldsymbol{B}_i \equiv \begin{bmatrix}\sigma^{-1}\boldsymbol{X}_i & \sigma^{-1}\boldsymbol{Z}'_i\\ \boldsymbol{O} & m^{-1/2}(\boldsymbol{I}_{m'}\otimes(\boldsymbol{\Sigma}')^{-1/2})\\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}\quad\text{and}\quad \overset{\bullet}{\boldsymbol{B}}_i \equiv \begin{bmatrix}\sigma^{-1}\boldsymbol{Z}_i\\ \boldsymbol{O}\\ \boldsymbol{\Sigma}^{-1/2}\end{bmatrix}.$$

Each of these matrices has $m'(n_{ii'}+q')+q$ rows. The $\boldsymbol{B}_i$ matrices each have $p+m'q'$ columns and the $\overset{\bullet}{\boldsymbol{B}}_i$ each have q columns. The solutions are

$$\widehat{\boldsymbol{\beta}} = \textit{first } p \textit{ rows of } \boldsymbol{x}_1,\ \ \text{Cov}(\widehat{\boldsymbol{\beta}}) = \textit{top left } p\times p \textit{ sub-block of } \boldsymbol{A}^{11},$$

$$\underset{1\le i'\le m'}{\text{stack}}\left(\widehat{\boldsymbol{u}}'_{i'}\right) = \textit{subsequent } (m'q')\times 1 \textit{ entries of } \boldsymbol{x}_1 \textit{ following } \widehat{\boldsymbol{\beta}},$$

$$E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}'_{i'}-\boldsymbol{u}'_{i'})^T\} = \textit{subsequent } p\times q' \textit{ sub-blocks of } \boldsymbol{A}^{11} \textit{ to the right of } \text{Cov}(\widehat{\boldsymbol{\beta}}),$$

$$\text{Cov}(\widehat{\boldsymbol{u}}'_{i'}-\boldsymbol{u}'_{i'}) = \textit{subsequent } q'\times q' \textit{ diagonal sub-blocks of } \boldsymbol{A}^{11} \textit{ following } \text{Cov}(\widehat{\boldsymbol{\beta}}),\ \ 1\le i'\le m',$$

$$\widehat{\boldsymbol{u}}_i = \boldsymbol{x}_{2,i},\ \ \text{Cov}(\widehat{\boldsymbol{u}}_i-\boldsymbol{u}_i) = \boldsymbol{A}^{22,i},\ \ E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i-\boldsymbol{u}_i)^T\} = \textit{first } p \textit{ rows of } \boldsymbol{A}^{12,i}$$

$$\underset{1\le i'\le m'}{\text{stack}}\left[E\{(\widehat{\boldsymbol{u}}_i-\boldsymbol{u}_i)(\widehat{\boldsymbol{u}}'_{i'}-\boldsymbol{u}'_{i'})^T\}\right] = \textit{remaining } m'q' \textit{ rows of } \boldsymbol{A}^{12,i},\ \ 1\le i\le m,$$

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).

Algorithm S.3 proceduralizes Result S.1 to facilitate computation of best linear unbiased predictors for the fixed and random effects parameters in (S.3) for fixed values of the covariance parameters. In practice, the covariance parameters would need to be replaced by estimates obtained using an approach such as restricted maximum likelihood. Algorithm S.3 also delivers the matrices in (S.5). In the case where m' is moderate but m is potentially very large Algorithm S.3 performs efficient streamlined computing.

Algorithm S.3 *Streamlined algorithm for obtaining best linear unbiased predictions and corresponding covariance matrix components for the linear mixed model with crossed random effects.*

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

Covariance Matrix Inputs: $\sigma^2 > 0, \ \ \boldsymbol{\Sigma}'(q' \times q'), \ \ \boldsymbol{\Sigma}(q \times q),$ symmetric and positive definite.

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

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{y}_i \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \ \boldsymbol{B}_i \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{X}_i & \sigma^{-1}\boldsymbol{Z}'_i \\ \boldsymbol{O} & m^{-1/2}(\boldsymbol{I}_{m'} \otimes (\boldsymbol{\Sigma}')^{-1/2}) \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix},$$

$$\overset{\bullet}{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_i \\ \boldsymbol{O} \\ \boldsymbol{\Sigma}^{-1/2} \end{bmatrix}$$

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

$\widehat{\boldsymbol{\beta}} \longleftarrow$ first p-rows of $\boldsymbol{x}_1$ component of $\mathcal{S}$

Cov$(\widehat{\boldsymbol{\beta}}) \longleftarrow$ top left $p \times p$ sub-block of $\boldsymbol{A}^{11}$ component of $\mathcal{S}$

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

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

  $i_{\text{end}} \longleftarrow i_{\text{stt}} + q' - 1$

  $\widehat{\boldsymbol{u}}_{i'} \longleftarrow$ sub-vector of $\boldsymbol{x}_1$ component of $\mathcal{S}$ with entries i_{stt} to i_{end}

  Cov$(\widehat{\boldsymbol{u}}'_{i'} - \boldsymbol{u}'_{i'}) \longleftarrow$ 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}

  $E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}'_{i'} - \boldsymbol{u}'_{i'})^T\} \longleftarrow$ sub-block of $\boldsymbol{A}^{11}$ component of $\mathcal{S}$ with rows 1 to p and columns i_{stt} to i_{end}

  $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$

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

  $\widehat{\boldsymbol{u}}_i \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}$; Cov$(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i) \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}$

  $E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)^T\} \longleftarrow$ sub-matrix of $\boldsymbol{A}^{12,i}$ component of $\mathcal{S}$ with rows 1 to p

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

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

    $i_{\text{end}} \longleftarrow i_{\text{stt}} + q' - 1$; $\boldsymbol{\Omega} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}$

    $E\{(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)(\widehat{\boldsymbol{u}}'_{i'} - \boldsymbol{u}'_{i'})^T\} \longleftarrow$ sub-matrix of $\boldsymbol{\Omega}^T$ with columns i_{stt} to i_{end}

    $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$

Outputs: $\widehat{\boldsymbol{\beta}}$, Cov$(\widehat{\boldsymbol{\beta}})$, $\Big\{ (\widehat{\boldsymbol{u}}'_{i'}, \, E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}'_{i'} - \boldsymbol{u}'_{i'})^T\}, \, \text{Cov}(\widehat{\boldsymbol{u}}'_{i'} - \boldsymbol{u}'_{i'})) : \ 1 \le i' \le m',$

  $\Big(\widehat{\boldsymbol{u}}_i, E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)^T\}, E\{(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)(\widehat{\boldsymbol{u}}'_{i'} - \boldsymbol{u}'_{i'})^T\}, \text{Cov}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i) \Big) : 1 \le i' \le m',$

  $1 \le i \le m \Big\}$

S.9 Full List of Items in the National Education Longitudinal Study

Table S.1 lists each of the 24 items within the National Education Longitudinal Study data set used in Section 6. Several of the measurements involve item response theory, which is abbreviated as IRT.

item	description
1	reading IRT-estimated number right
2	mathematics IRT-estimated number right
3	science IRT-estimated number right
4	history/citizenship/geography IRT-estimated number right
5	reading standardized score
6	mathematics standardized score
7	science standardized score
8	history/citizenship/geography standardized score
9	reading IRT estimate of ability
10	mathematics IRT estimate of ability
11	science IRT estimate of ability
12	history/citizenship/geography IRT estimate of ability
13	standardized test composite (reading, mathematics)
14	reading level 1: probability of proficiency
15	reading level 2: probability of proficiency
16	reading level 3: probability of proficiency
17	mathematics level 1: probability of proficiency
18	mathematics level 2: probability of proficiency
19	mathematics level 3: probability of proficiency
20	mathematics level 4: probability of proficiency
21	science level 1: probability of proficiency
22	science level 2: probability of proficiency
23	science level 3: probability of proficiency
24	science level 4: probability of proficiency

Table S.1: *Descriptions of each of the 24 items in the National Education Longitudinal Study data used in Section 6. The abbreviation IRT stands for item response theory. Fuller details are Thurgood et al. (2003).*