Streamlined Computing for Variational Inference
with Higher Level Random Effects

BY TUI H. NOLAN, MARIANNE MENICTAS AND MATT P. WAND

University of Technology Sydney

12th March, 2019

Abstract

We derive and present explicit algorithms to facilitate streamlined computing for variational inference for models containing higher level random effects. Existing literature, such as Lee & Wand (2016), is such that streamlined variational inference is restricted to mean field variational Bayes algorithms for two-level random effects models. Here we provide the following extensions: (1) explicit Gaussian response mean field variational Bayes algorithms for three-level models, (2) explicit algorithms for the alternative variational message passing approach in the case of two-level and three-level models, and (3) an explanation of how arbitrarily high levels of nesting can be handled based on the recently published matrix algebraic results of the authors. A pay-off from (2) is simple extension to non-Gaussian response models. In summary, we remove barriers for streamlining variational inference algorithms based on either the mean field variational Bayes approach or the variational message passing approach when higher level random effects are present.

Keywords: Factor graph fragment; Longitudinal data analysis; Mixed models; Multilevel models; Variational message passing.

1 Introduction

Models involving higher level random effects commonly arise in a variety of contexts. The areas of study known as longitudinal data analysis (e.g. Fitzmaurice *et al.*, 2008), mixed models (e.g. Pinheiro & Bates, 2000), multilevel models (e.g. Goldstein, 2010), panel data analysis (e.g. Baltagi, 2013) and small area estimation (e.g. Rao & Molina, 2015) potentially each require the handling of higher levels of nesting. Our main focus in this article is providing explicit algorithms that facilitate variational inference for up to three-level random effects and a pathway for handling even higher levels. Both direct and message passing approaches to mean field variational Bayes are treated. We also provide algorithms for frequentist best linear unbiased prediction that, to our knowledge, are not in the current literature and may be viewed of extensions of the exquisite streamlined mixed model computational strategies presented in Section 2.2 of Pinheiro & Bates (2000).

A useful prototype setting for understanding the nature and computational challenges is a fictitious sociology example in which residents (level 1 units) are divided into different towns (level 2 units) and those towns are divided into different districts (level 3 units). Following Goldstein (2010) we call these three-level data, although note that Pinheiro & Bates (2000) use the term "two-level", corresponding to two levels of nesting, for the same setting. Figure 1 displays simulated regression data generated according to this setting with a single predictor variable corresponding to years of education and the response corresponding to annual income. In Figure 1, the number of districts is 6, the number of towns per district is 8 and the resident sample size within each town is 25. In each panel of Figure 1, the line corresponds to the best linear unbiased prediction fit of a three-level random intercepts and slopes linear mixed model, as explained in Section 5.1. The variational Bayesian analogue, covered in Section 5.2, is such that best linear unbiased prediction is re-
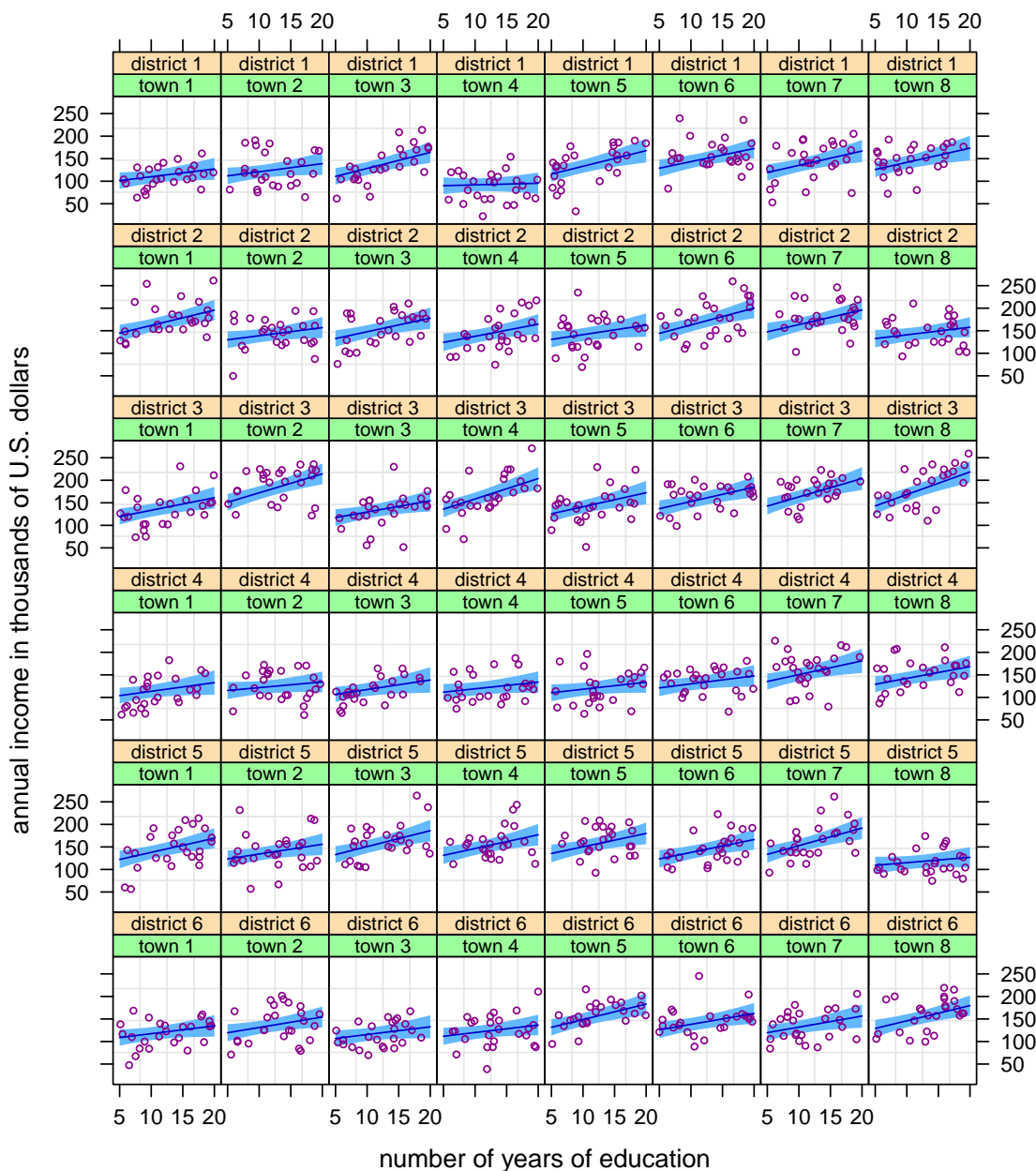
Figure 1: *Simulated three-level data according to 6 districts, each having 8 towns, each having 25 randomly chosen residents. In each panel, the line corresponds to a best linear unbiased prediction fit, according to an appropriate multilevel model and the shaded region corresponds to pointwise 95% confidence intervals for the mean response.*

placed by variational approximate posterior means and confidence intervals are replaced by variational approximate credible intervals. Now suppose that the group and sample sizes are much larger with, say, 500 districts, 60 towns per district and 1,000 residents per town. Then naïve fitting is storage-greedy and computationally challenging since the combined fixed and random effects design matrices have 1.83×10^{12} entries of which at least 99.99% equal zero. A major contribution of this article is explaining how variational inference can be achieved using only the 0.01% non-zero design matrix components with updates that are linear in the numbers of groups.

Our streamlined variational inference algorithms for higher level random effects models rely on four theorems provided by Nolan & Wand (2019) concerning linear system

solutions and sub-blocks of matrix inverses for two-level and three-level sparse matrix problems which are the basis for the fundamental Algorithms 1–4 in Section 3. In that article, as well as here, we treat one higher level situation at a time. Even though four-level and even higher level situations may be of interest in future analysis, the required theory is not yet in place. As we will see, covering both direct and message passing approaches for just the two-level and three-level cases is quite a big task. Nevertheless, our results and algorithms shed important light on streamlined variational inference for general higher level random effects models.

After laying out the four fundamental algorithms in Section 3 we then derive an additional ten algorithms, labeled Algorithms 5–14, that facilitate streamlined frequentist and variational inference for two-level and three-level linear mixed models. Algorithms 5 and 10 treat best linear unbiased prediction as a prelude to the closely related mean field variational Bayes analogues, which are dealt with in Algorithms 6 and 11. The remaining four algorithms are concerned with streamlined factor graph fragment updates according to the variational message passing infrastructure described in Wand (2017). As explained there, the message passing approach has the advantage compartmentalization of variational inference algebra and code. The inherent complexity of streamlined variational inference for higher level random effects models is such that the current article is restricted to ordinary linear mixed models. Extensions such as generalized additive mixed models with higher level random effects and higher level group-specific curve models follow from Algorithms 1–4, but must be treated elsewhere. Section 8 provides further details on this matter.

Our algorithms also build on previous work on streamlined variational inference for similar classes of models described in Lee & Wand (2016). However, Lee & Wand (2016) only treated the two-level case, did not employ QR decomposition enhancement and did not include any variational message passing algorithms. The current article is a systematic treatment of higher level random effects models beyond the common two-level case.

Section 2 provides background material concerning variational inference. In Section 3 we present four algorithms for solving higher level sparse matrix problems which are fundamental for variational inference involving general models with hierarchical random effects structure. The two-level situation is treated in Section 4, followed by treatment of the three-level situation in Section 5. Section 6 demonstrates the speed advantages of streamlining for variational inference in random effects models via some timing studies. Illustration for data from a large perinatal health study is given in Section 7. In Section 8 we close with some discussion about extensions to other settings.

2 Variational Inference Background

In keeping with the theme of this article, we will explain the essence of variational inference for a general class of Bayesian linear mixed models. Summaries of variational inference in wider statistical contexts are given in Ormerod & Wand (2010) and Blei, Kucukelbir & McAuliffe (2017).

Suppose that the response data vector $\boldsymbol{y}$ is modeled according to a Bayesian version of the Gaussian linear mixed model (e.g. Robinson, 1991)

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{R} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \boldsymbol{R}), \quad \boldsymbol{u}|\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}) \tag{1}$$

for hyperparameters $\boldsymbol{\mu_\beta}$ and $\boldsymbol{\Sigma_\beta}$ and such that $\boldsymbol{\beta}$ and $\boldsymbol{u}|\boldsymbol{G}$ are independent. The $\boldsymbol{\beta}$ and $\boldsymbol{u}$ vectors are labeled *fixed effects* and *random effects*, respectively. Their corresponding *design matrices* are $\boldsymbol{X}$ and $\boldsymbol{Z}$. We will allow for the possibility that prior specification for the covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ involves auxiliary covariance matrices $\boldsymbol{A_G}$ and $\boldsymbol{A_R}$ with conjugate Inverse G-Wishart distributions (Wand, 2017). The prior specification of $\boldsymbol{G}$ and $\boldsymbol{R}$ involves the specifications

$$\mathfrak{p}(\boldsymbol{G}|\boldsymbol{A_G}), \quad \mathfrak{p}(\boldsymbol{A_G}), \quad \mathfrak{p}(\boldsymbol{R}|\boldsymbol{A_R}) \quad \text{and} \quad \mathfrak{p}(\boldsymbol{A_R}). \tag{2}$$

Figure 2 is a directed acyclic graph representation of (1) and (2). The circles, usually called *nodes*, correspond to the model's random vectors and random matrices. The arrows depict conditional independence relationships (e.g. Bishop, 2006; Chapter 8).
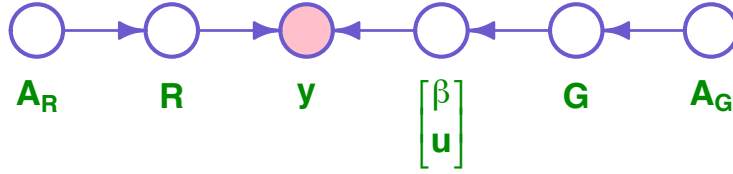

Figure 2: *Directed acyclic graph representation of model (1). The shading of the $\boldsymbol{y}$ node indicates that this vector of response values is observed.*

Full Bayesian inference for the $\boldsymbol{\beta}$, $\boldsymbol{G}$ and $\boldsymbol{R}$ and the random effects $\boldsymbol{u}$ involves the posterior density function $\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_G, \boldsymbol{A}_R, \boldsymbol{G}, \boldsymbol{R}|\boldsymbol{y})$, but typically is analytically intractable and Markov chain Monte Carlo approaches are required for practical 'exact' inference. Variational approximate inference involves mean field restrictions such as

$$\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_G, \boldsymbol{A}_R, \boldsymbol{G}, \boldsymbol{R}|\boldsymbol{y}) \approx \mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_G, \boldsymbol{A}_R)\,\mathfrak{q}(\boldsymbol{G}, \boldsymbol{R}) \tag{3}$$

for density functions $\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_G, \boldsymbol{A}_R)$ and $\mathfrak{q}(\boldsymbol{G}, \boldsymbol{R})$, which we call $\mathfrak{q}$-*densities*. The approximation at (3) represents the minimal product restriction for which practical variational inference algorithms arise. However, as explained in Section 10.2.5 of Bishop (2006), the graphical structure of Figure 2 induces further product density forms and the right-hand side of (3) admits the further factorization

$$\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})\mathfrak{q}(\boldsymbol{A}_G)\mathfrak{q}(\boldsymbol{A}_R)\mathfrak{q}(\boldsymbol{G})\mathfrak{q}(\boldsymbol{R}). \tag{4}$$

With this product density form in place, the forms and optimal parameters for the $\mathfrak{q}$-densities are obtained by minimising the Kullback-Leibler divergence of the right-hand side of (3) from its left-hand side. The optimal $\mathfrak{q}$-density parameters are interdependent and a coordinate ascent algorithm (e.g. Algorithm 1 of Ormerod & Wand, 2010) is used to obtain their solution. For example, the optimal $\mathfrak{q}$-density for $(\boldsymbol{\beta}, \boldsymbol{u})$, denoted by $\mathfrak{q}^*(\boldsymbol{\beta}, \boldsymbol{u})$, is a Multivariate Normal density function with mean vector $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$. The coordinate ascent algorithm is such that they are updated according to

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})} \longleftarrow \left\{ \boldsymbol{C}^T E_{\mathfrak{q}}(\boldsymbol{R}^{-1})\boldsymbol{C} + \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & E_{\mathfrak{q}}(\boldsymbol{G}^{-1}) \end{bmatrix} \right\}^{-1}$$

$$\text{and} \quad \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})} \longleftarrow \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}\boldsymbol{C}^T E_{\mathfrak{q}}(\boldsymbol{R}^{-1})\left(\boldsymbol{y} + \begin{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix} \right) \tag{5}$$

where $E_{\mathfrak{q}}(\boldsymbol{G}^{-1})$ and $E_{\mathfrak{q}}(\boldsymbol{R}^{-1})$ are the $\mathfrak{q}$-density expectations of $\boldsymbol{G}^{-1}$ and $\boldsymbol{R}^{-1}$ and $\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}]$. If, for example, (1) corresponds to a mixed model with three-level random effects such that $\boldsymbol{R} = \sigma^2 \boldsymbol{I}$ then, as pointed out in Section 1, with 60 groups at level 2 and 500 groups at level 3 the matrix $\boldsymbol{C}$ has almost 2 trillion entries of which 99.99% are zero. Moreover, $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ is a $61,002 \times 61,002$ matrix of which only about 0.016% of its approximately 3.7 billion entries are required for variational inference under mean field restriction (3). Avoiding the wastage of the naïve updates given by (5) is the crux of this article and dealt with in the upcoming sections. The updates for $E_{\mathfrak{q}}(\boldsymbol{G}^{-1})$ and $E_{\mathfrak{q}}(\boldsymbol{R}^{-1})$ depend on parameterizations of $\boldsymbol{G}$ and $\boldsymbol{R}$. For example, $\boldsymbol{R} = \sigma^2\boldsymbol{I}$ for some $\sigma^2 > 0$ throughout Sections 4 and 5. However, these covariance parameter updates are relatively simple and free of storage and computational efficiency issues. Similar comments apply to the updates for the $\mathfrak{q}$-density parameters of $\boldsymbol{A}_G$ and $\boldsymbol{A}_R$.

An alternative approach to obtaining $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ the relevant sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and the covariance and auxiliary variable q-parameter updates is to use the notion of *message passing* on a *factor graph*. The relevant factor graph for model (1), according to the product density form (4), is shown in Figure 3.
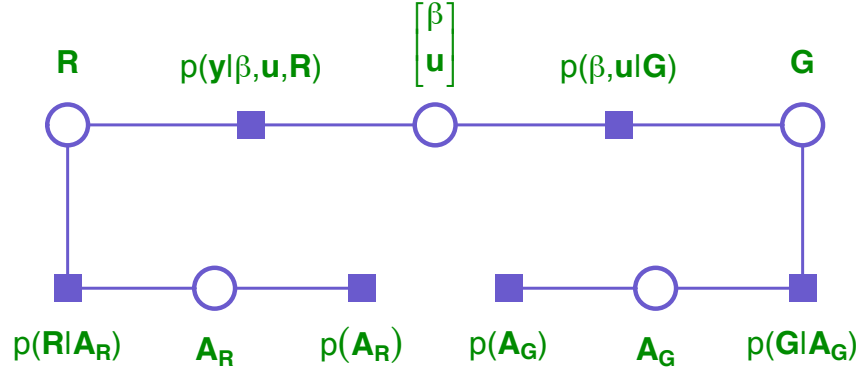

Figure 3: *Factor graph representation of the product structure of (6) with the solid rectangles corresponding to the factors and open circles corresponding to the unobserved random vectors and random matrices of the Bayesian linear mixed model given by (1) and (2), known as stochastic nodes. Edges join each factor to the stochastic nodes that are present in the factor.*

The circles in Figure 3 correspond to the parameters in each factor of (4) and are referred to as *stochastic nodes*. The squares correspond to the factors of

$$\mathfrak{p}(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{A}_G,\boldsymbol{A}_R,\boldsymbol{G},\boldsymbol{R}) = \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{R})\,\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{G})\,\mathfrak{p}(\boldsymbol{G}|\boldsymbol{A}_G)\,\mathfrak{p}(\boldsymbol{R}|\boldsymbol{A}_R)\,\mathfrak{p}(\boldsymbol{A}_G)\,\mathfrak{p}(\boldsymbol{A}_R), \quad (6)$$

with factorization according to the conditional independence structure apparent from Figure 2. Then, as explained in e.g. Minka (2005), the $\mathfrak{q}$-density of $(\boldsymbol{\beta},\boldsymbol{u})$ can be expressed as

$$\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}) \propto m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{R}) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\; m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{G}) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})$$

where

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{R}) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \quad \text{and} \quad m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{G}) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})$$

are known as *messages*, with the subscripts indicating that they are passed from $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{R})$ to $(\boldsymbol{\beta},\boldsymbol{u})$ and $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{G})$ to $(\boldsymbol{\beta},\boldsymbol{u})$ respectively. Messages are simply functions of the stochastic node to which the message is passed and, for mean field variational inference, are formed according to rules listed in Minka (2005) and Section 2.5 of Wand (2017). To compartmentalize algebra and coding for variational message passing, Wand (2017) advocates the use of *fragments*, which are sub-graphs of a factor graph containing a single factor and each of its neighboring stochastic nodes. In Sections 4 and 5 of Wand (2017), eight important fragments are identified and treated including those needed for a wide range of linear mixed models. However, in the interests of brevity, Wand (2017) ignored issues surrounding potentially very large and sparse matrices in the message parameter vectors. In Sections 4 and 5 of this article, we explain how the messages passed to the $(\boldsymbol{\beta},\boldsymbol{u})$ node can be streamlined to avoid massive sparse matrices.

A core component of the message passing approach to variational inference is exponential family forms, sufficient statistics and natural parameters. For a $d \times 1$ Multivariate Normal random vector

$$\boldsymbol{x} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

this involves re-expression of its density function according to

$$\mathfrak{p}(\boldsymbol{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

$$= \exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta} - A(\boldsymbol{\eta}) - \tfrac{d}{2}\log(2\pi)\}$$

where

$$T(x) \equiv \left[\begin{array}{c} x \\ \text{vech}(x\,x^T) \end{array} \right] \quad \text{and} \quad \eta \equiv \left[\begin{array}{c} \eta_1 \\ \eta_2 \end{array} \right] \equiv \left[\begin{array}{c} \Sigma^{-1}\mu \\ -\frac{1}{2}\,D_d^T\text{vec}(\Sigma^{-1}) \end{array} \right]$$

are, respectively, the *sufficient statistic* and *natural parameter* vectors. The matrix D_d, known as the *duplication matrix of order d*, is the $d^2 \times \{\frac{1}{2}d(d+1)\}$ matrix containing only zeroes and ones such that $D_d\,\text{vech}(A) = \text{vec}(A)$ for any symmetric $d \times d$ matrix A. The function

$$A(\eta) = -\tfrac{1}{4}\,\eta_1^T \left\{ \text{vec}^{-1}(D_d^{+T}\eta_2) \right\}^{-1} \eta_1 - \tfrac{1}{2}\log\left|-2\,\text{vec}^{-1}(D_d^{+T}\eta_2)\right|$$

is the *log-partition* function, where $D_d^+ \equiv (D_d^T D_d)^{-1} D_d^T$ is the Moore-Penrose inverse of D_d and is such that $D_d^+\text{vech}(A) = \text{vec}(A)$ whenever A is symmetric. The inverse of the natural parameter transformation is given by

$$\mu = -\tfrac{1}{2}\left\{\text{vec}^{-1}(D_d^{+T}\eta_2)\right\}^{-1}\eta_1 \quad \text{and} \quad \Sigma = -\tfrac{1}{2}\left\{\text{vec}^{-1}(D_d^{+T}\eta_2)\right\}^{-1}.$$

The vec and vech matrix operators are reasonably well-established (e.g. Gentle, 2007). If a is a $d^2 \times 1$ vector then $\text{vec}^{-1}(a)$ is the $d \times d$ matrix such that $\text{vec}(A) = a$. We also require vec inversion to non-square matrices. If a is a $(d_1 d_2) \times 1$ vector then $\text{vec}^{-1}_{d_1 \times d_2}(a)$ is the $d_1 \times d_2$ matrix such that $\text{vec}(A) = a$.

The other major distributional family used throughout this article is a generalization of the Inverse Wishart distribution known as the *Inverse G-Wishart* distribution. It 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 M *respects G* if

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

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

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

if and only if the density function of X satisfies

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

over arguments X such that X is symmetric and positive definite and 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 X^{-1} is a full matrix and 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 message passing 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. Throughout this article 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.

Finally, we remark on the $\mathfrak{p}$ and $\mathfrak{q}$ notation used for density functions in this article. In the variational inference literature these letters have become very commonplace to denote the density functions corresponding to the model and the density functions of parameters according to the mean field approximation, with $\mathfrak{p}$ for the former and $\mathfrak{q}$ for the latter. However, the same letters are commonly used as dimension variables in the mixed models literature (e.g. Pinheiro & Bates, 2000). Therefore we use ordinary p and q as dimension variables and scripted versions of these letters ($\mathfrak{p}$ and $\mathfrak{q}$) for density functions.

3 Multilevel Sparse Matrix Problem Algorithms

A key observation in this work is the fact that streamlining of variational inference algorithms for higher level random effects models can be achieved by recognition and isolation of a few fundamental algorithms, which we call multilevel sparse matrix problem algorithms. These algorithms, based on the results of Nolan & Wand (2019), are identical to those used traditionally for fitting frequentist random effects (Pinheiro & Bates, 2000). For each level there are two types of sparse matrix solution algorithms: one that applies to general forms and one that uses a QR-decomposition enhancement for a particular form that arises commonly for models containing random effects. Both types are needed for variational inference.

In theory, based on the infrastructure laid out in Nolan & Wand (2019), any number of levels can be handled. However, each higher level brings increasing complexity. Here we restrict attention to two-level and three-level sparse matrix algorithms.

3.1 Two-Level Sparse Matrix Algorithms

Two-level sparse matrix problems are described in Section 2 of Nolan & Wand (2019). The notation used there is also used in this section. Here we present two algorithms, named

$$\textsc{SolveTwoLevelSparseMatrix} \quad \text{and} \quad \textsc{SolveTwoLevelSparseLeastSquares}$$

which are at the heart of streamlining variational inference for two-level models.

The SolveTwoLevelSparseMatrix algorithm is concerned with solving general two-level sparse linear system problem $\boldsymbol{Ax} = \boldsymbol{a}$, where

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \cdots & \boldsymbol{A}_{12,m} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{A}_{12,m}^T & \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{A}_{22,m} \end{bmatrix}, \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \vdots \\ \boldsymbol{a}_{2,m} \end{bmatrix} \quad \text{and} \quad \boldsymbol{x} \equiv \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_{2,1} \\ \boldsymbol{x}_{2,2} \\ \vdots \\ \boldsymbol{x}_{2,m} \end{bmatrix} \quad (7)$$

and obtaining the sub-matrices corresponding to the non-zero blocks of $\boldsymbol{A}$:

$$
\boldsymbol{A}^{-1} \equiv
\begin{bmatrix}
\boldsymbol{A}^{11} & \boldsymbol{A}^{12,1} & \boldsymbol{A}^{12,2} & \cdots & \boldsymbol{A}^{12,m} \\
\hline
\boldsymbol{A}^{12,1\,T} & \boldsymbol{A}^{22,1} & \times & \cdots & \times \\
\hline
\boldsymbol{A}^{12,2\,T} & \times & \boldsymbol{A}^{22,2} & \cdots & \times \\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots \\
\hline
\boldsymbol{A}^{12,m\,T} & \times & \times & \cdots & \boldsymbol{A}^{22,m}
\end{bmatrix}.
\tag{8}
$$

As will be elaborated upon later, the blocks represented by the $\times$ symbol are not of interest. SOLVETWOLEVELSPARSEMATRIX is listed as Algorithm 1 and is justified by Theorem 1 of Nolan & Wand (2019).

Algorithm 1 *The* SOLVETWOLEVELSPARSEMATRIX *algorithm for solving the two-level sparse matrix problem* $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{a}$ *and sub-blocks of* $\boldsymbol{A}^{-1}$ *corresponding to the non-zero sub-blocks of* $\boldsymbol{A}$. *The sub-block notation is given by (7) and (8).*

Inputs: $\Big(\boldsymbol{a}_1(p \times 1), \boldsymbol{A}_{11}(p \times p), \big\{\big(\boldsymbol{a}_{2,i}(q \times 1), \boldsymbol{A}_{22,i}(q \times q), \boldsymbol{A}_{12,i}(p \times q)\big) : 1 \le i \le m\big\}$

$\boldsymbol{\omega}_1 \longleftarrow \boldsymbol{a}_1$; $\boldsymbol{\Omega}_2 \longleftarrow \boldsymbol{A}_{11}$

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

$\quad \boldsymbol{\omega}_1 \longleftarrow \boldsymbol{\omega}_1 - \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{a}_{2,i}$; $\boldsymbol{\Omega}_2 \longleftarrow \boldsymbol{\Omega}_2 - \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^T$

$\boldsymbol{A}^{11} \longleftarrow \boldsymbol{\Omega}_2^{-1}$; $\boldsymbol{x}_1 \longleftarrow \boldsymbol{A}^{11}\boldsymbol{\omega}_1$

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

$\quad \boldsymbol{x}_{2,i} \longleftarrow \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{a}_{2,i} - \boldsymbol{A}_{12,i}^T\boldsymbol{x}_1)$; $\boldsymbol{A}^{12,i} \longleftarrow -(\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^T\boldsymbol{A}^{11})^T$

$\quad \boldsymbol{A}^{22,i} \longleftarrow \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{I} - \boldsymbol{A}_{12,i}^T\boldsymbol{A}^{12,i})$

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

The SOLVETWOLEVELSPARSELEASTSQUARES algorithm arises in the special case where $\boldsymbol{x}$ is the minimizer of the least squares problem $\|\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x}\|^2 \equiv (\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x})^T(\boldsymbol{b} - \boldsymbol{B}\boldsymbol{x})$ where

$$
\boldsymbol{B} \equiv
\begin{bmatrix}
\boldsymbol{B}_1 & \dot{\boldsymbol{B}}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\
\hline
\boldsymbol{B}_2 & \boldsymbol{O} & \dot{\boldsymbol{B}}_2 & \cdots & \boldsymbol{O} \\
\hline
\vdots & \vdots & \vdots & \ddots & \vdots \\
\hline
\boldsymbol{B}_m & \boldsymbol{O} & \boldsymbol{O} & \cdots & \dot{\boldsymbol{B}}_m
\end{bmatrix}
\quad \text{and} \quad
\boldsymbol{b} \equiv
\begin{bmatrix}
\boldsymbol{b}_1 \\
\hline
\boldsymbol{b}_2 \\
\hline
\vdots \\
\hline
\boldsymbol{b}_m
\end{bmatrix}.
\tag{9}
$$

In this case $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$, $\boldsymbol{a} = \boldsymbol{B}^T\boldsymbol{b}$ so that the sub-blocks of $\boldsymbol{A}$ and $\boldsymbol{a}$ take the forms

$$
\boldsymbol{A}_{11} = \sum_{i=1}^m \boldsymbol{B}_i^T\boldsymbol{B}_i, \quad \boldsymbol{A}_{12,i} = \boldsymbol{B}_i^T\dot{\boldsymbol{B}}_i, \quad \boldsymbol{A}_{22,i} = \dot{\boldsymbol{B}}_i^T\dot{\boldsymbol{B}}_i, \quad \boldsymbol{a}_1 = \sum_{i=1}^m \boldsymbol{B}_i^T\boldsymbol{b}_i \quad \text{and} \quad \boldsymbol{a}_{2,i} = \dot{\boldsymbol{B}}_i^T\boldsymbol{b}_i.
$$

As demonstrated in Section 4, these forms arise in two-level random effects models. Theorem 2 of Nolan & Wand (2019) shows that this special form lends itself to a QR decomposition (e.g. Harville, 2008; Section 6.4.d) approach which has speed and stability advantages in regression settings (e.g. Gentle, 2007; Section 6.7.2).

SOLVETWOLEVELSPARSELEASTSQUARES is listed as Algorithm 2. Note that we use $\tilde{n}_i$, rather than n_i, to denote the number of rows in each of $\boldsymbol{b}_i$, $\boldsymbol{B}_i$ and $\dot{\boldsymbol{B}}_i$ to avoid a notational clash with common grouped data dimension notation as used in Section 4. In the first loop over the m groups of data the upper triangular matrices $\boldsymbol{R}_i$, $1 \le i \le m$, are obtained via QR-decomposition; a standard procedure within most computing environments. Following that, all matrix equations involve $\boldsymbol{R}_i^{-1}$, which can be achieved rapidly via back-solving.

Algorithm 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 (7), (8) and (9).*

Input: $\left\{ \left(\boldsymbol{b}_i(\tilde{n}_i \times 1),\ \boldsymbol{B}_i(\tilde{n}_i \times p),\ \dot{\boldsymbol{B}}_i(\tilde{n}_i \times q) \right) : 1 \le i \le m \right\}$

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

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

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

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

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

$\quad$ $\boldsymbol{C}_{1i} \longleftarrow$ 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 $\boldsymbol{\Omega}_4 = \boldsymbol{Q} \begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix}$ 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 $i = 1, \ldots, m$:

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

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

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

Note that in Algorithm 2 calculations such as $\boldsymbol{Q}_i^T \boldsymbol{B}_i$ do not require storage of $\boldsymbol{Q}_i$ and use of ordinary multiplication. Standard matrix algebraic programming languages store information concerning $\boldsymbol{Q}_i$ in a compact form from which matrices such as $\boldsymbol{Q}_i^T \boldsymbol{B}_i$ can be efficiently obtained.

3.2 Three-Level Sparse Matrix Algorithms

Extension to the three-level situation is described in Section 3 of Nolan & Wand (2019). Theorems 3 and 4 given there lead to the algorithms

$\quad$ SOLVETHREELEVELSPARSEMATRIX and SOLVETHREELEVELSPARSELEASTSQUARES

which facilitate streamlining variational inference for three-level models.

An illustrative three-level sparse matrix is:

$$
\boldsymbol{A} =
\left[
\begin{array}{c|c|c|c|c|c|c|c}
\boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,11} & \boldsymbol{A}_{12,12} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,21} & \boldsymbol{A}_{12,22} & \boldsymbol{A}_{12,23} \\ \hline
\boldsymbol{A}_{12,1}^{T} & \boldsymbol{A}_{22,1} & \boldsymbol{A}_{12,1,1} & \boldsymbol{A}_{12,1,2} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \hline
\boldsymbol{A}_{12,11}^{T} & \boldsymbol{A}_{12,1,1}^{T} & \boldsymbol{A}_{22,11} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \hline
\boldsymbol{A}_{12,12}^{T} & \boldsymbol{A}_{12,1,2}^{T} & \boldsymbol{O} & \boldsymbol{A}_{22,12} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \hline
\boldsymbol{A}_{12,2}^{T} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{A}_{12,2,1} & \boldsymbol{A}_{12,2,2} & \boldsymbol{A}_{12,2,3} \\ \hline
\boldsymbol{A}_{12,21}^{T} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{12,2,1}^{T} & \boldsymbol{A}_{22,21} & \boldsymbol{O} & \boldsymbol{O} \\ \hline
\boldsymbol{A}_{12,22}^{T} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{12,2,2}^{T} & \boldsymbol{O} & \boldsymbol{A}_{22,22} & \boldsymbol{O} \\ \hline
\boldsymbol{A}_{12,23}^{T} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{12,2,3}^{T} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,23}
\end{array}
\right]
\tag{10}
$$

and corresponds to level 2 group sizes of $n_1 = 2$ and $n_2 = 3$, and a level 3 group size of $m = 2$. A general three-level sparse matrix $\boldsymbol{A}$ consists of the following components:

- A $p \times p$ matrix $\boldsymbol{A}_{11}$, which is designated the $(1,1)$-block position.

- A set of partitioned matrices $\left\{ \left[\ \boldsymbol{A}_{12,i} \ \middle| \ \boldsymbol{A}_{12,ij} \ \middle| \ \ldots \ \middle| \ \boldsymbol{A}_{12,in_i} \ \right] : 1 \leq i \leq m \right\}$, which is designated the $(1,2)$-block position. For each $1 \leq i \leq m$, $\boldsymbol{A}_{12,i}$ is $p \times q_1$, and for each $1 \leq j \leq n_i$, $\boldsymbol{A}_{12,ij}$ is $p \times q_2$.

- A $(2,1)$-block, which is simply the transpose of the $(1,2)$-block.

- A block diagonal structure along the $(2,2)$-block position, where each sub-block is a two-level sparse matrix, as defined in (7). For each $1 \leq i \leq m$, $\boldsymbol{A}_{22,i}$ is $q_1 \times q_1$, and for each $1 \leq j \leq n_i$, $\boldsymbol{A}_{12,i,j}$ is $q_1 \times q_2$ and $\boldsymbol{A}_{22,ij}$ is $q_2 \times q_2$.

The three-level sparse linear system problem takes the form $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{a}$ where we partition the vectors $\boldsymbol{a}$ and $\boldsymbol{x}$ as follows:

$$
\boldsymbol{a} \equiv
\left[
\begin{array}{c}
\boldsymbol{a}_1 \\ \hline
\boldsymbol{a}_{2,1} \\ \hline
\boldsymbol{a}_{2,11} \\ \hline
\boldsymbol{a}_{2,12} \\ \hline
\boldsymbol{a}_{2,2} \\ \hline
\boldsymbol{a}_{2,21} \\ \hline
\boldsymbol{a}_{2,22} \\ \hline
\boldsymbol{a}_{2,23}
\end{array}
\right]
\quad \text{and} \quad
\boldsymbol{x} \equiv
\left[
\begin{array}{c}
\boldsymbol{x}_1 \\ \hline
\boldsymbol{x}_{2,1} \\ \hline
\boldsymbol{x}_{2,11} \\ \hline
\boldsymbol{x}_{2,12} \\ \hline
\boldsymbol{x}_{2,2} \\ \hline
\boldsymbol{x}_{2,21} \\ \hline
\boldsymbol{x}_{2,22} \\ \hline
\boldsymbol{x}_{2,23}
\end{array}
\right].
\tag{11}
$$

Here $\boldsymbol{a}_1$ and $\boldsymbol{x}_1$ are $p \times 1$ vectors. Then, for each $1 \leq i \leq m$, $\boldsymbol{a}_{2,i}$ and $\boldsymbol{x}_{2,i}$ are $q_1 \times 1$ vectors. Lastly, for each $1 \leq i \leq m$ and $1 \leq j \leq n_i$ the vectors $\boldsymbol{a}_{2,ij}$ and $\boldsymbol{x}_{2,ij}$ have dimension $q_2 \times 1$.

The three-level sparse matrix inverse problem involves determination of the sub-blocks of $\boldsymbol{A}^{-1}$ corresponding to the non-zero sub-blocks of $\boldsymbol{A}$. Our notation for these sub-blocks

is illustrated by

$$
\boldsymbol{A}^{-1} =
\begin{bmatrix}
\boldsymbol{A}^{11} & \boldsymbol{A}^{12,1} & \boldsymbol{A}^{12,11} & \boldsymbol{A}^{12,12} & \boldsymbol{A}^{12,2} & \boldsymbol{A}^{12,21} & \boldsymbol{A}^{12,22} & \boldsymbol{A}^{12,23} \\
\boldsymbol{A}^{12,1\,T} & \boldsymbol{A}^{22,1} & \boldsymbol{A}^{12,1,1} & \boldsymbol{A}^{12,1,2} & \times & \times & \times & \times \\
\boldsymbol{A}^{12,11\,T} & \boldsymbol{A}^{12,1,1\,T} & \boldsymbol{A}^{22,11} & \times & \times & \times & \times & \times \\
\boldsymbol{A}^{12,12\,T} & \boldsymbol{A}^{12,1,2\,T} & \times & \boldsymbol{A}^{22,12} & \times & \times & \times & \times \\
\boldsymbol{A}^{12,2\,T} & \times & \times & \times & \boldsymbol{A}^{22,2} & \boldsymbol{A}^{12,2,1} & \boldsymbol{A}^{12,2,2} & \boldsymbol{A}^{12,2,3} \\
\boldsymbol{A}^{12,21\,T} & \times & \times & \times & \boldsymbol{A}^{12,2,1\,T} & \boldsymbol{A}^{22,21} & \times & \times \\
\boldsymbol{A}^{12,22\,T} & \times & \times & \times & \boldsymbol{A}^{12,2,2\,T} & \times & \boldsymbol{A}^{22,22} & \times \\
\boldsymbol{A}^{12,23\,T} & \times & \times & \times & \boldsymbol{A}^{12,2,3\,T} & \times & \times & \boldsymbol{A}^{22,23}
\end{bmatrix}
\tag{12}
$$

for the $m = 2$, $n_1 = 2$ and $n_2 = 3$ case.

SOLVETHREELEVELSPARSEMATRIX, which provides streamlined solutions for the general three-level sparse matrix problem, is listed as Algorithm 3.

Algorithm 3 *The* SOLVETHREELEVELSPARSEMATRIX *algorithm for solving the three-level sparse matrix problem $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{a}$ and sub-blocks of $\boldsymbol{A}^{-1}$ corresponding to the non-zero sub-blocks of $\boldsymbol{A}$. The sub-block notation is given by (10), (11) and (12).*

Input: $\Big(\boldsymbol{a}_1(p \times 1),\, \boldsymbol{A}_{11}(p \times p),\, \big\{\big(\boldsymbol{a}_{2,i}(q_1 \times 1),\, \boldsymbol{A}_{22,i}(q_1 \times q_1),\, \boldsymbol{A}_{12,i}(p \times q_1) :\ 1 \le i \le m\big\},$

$\big\{\boldsymbol{a}_{2,ij}(q_2 \times 1),\, \boldsymbol{A}_{22,ij}(q_2 \times q_2),\, \boldsymbol{A}_{12,ij}(p \times q_2),\, \boldsymbol{A}_{12,\,i,\,j}(q_1 \times q_2)\big) :\ 1 \le i \le m,\ 1 \le j \le n_i\big\}\Big).$

$\boldsymbol{\omega}_5 \longleftarrow \boldsymbol{a}_1$; $\boldsymbol{\Omega}_6 \longleftarrow \boldsymbol{A}_{11}$

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

$\qquad \boldsymbol{h}_{2,i} \longleftarrow \boldsymbol{a}_{2,i}$; $\boldsymbol{H}_{12,i} \longleftarrow \boldsymbol{A}_{12,i}$; $\boldsymbol{H}_{22,i} \longleftarrow \boldsymbol{A}_{22,i}$

$\qquad$ For $j = 1, \ldots, n_i$:

$\qquad\qquad \boldsymbol{h}_{2,i} \longleftarrow \boldsymbol{h}_{2,i} - \boldsymbol{A}_{12,\,i,\,j}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{a}_{2,ij}$; $\boldsymbol{H}_{12,i} \longleftarrow \boldsymbol{H}_{12,i} - \boldsymbol{A}_{12,ij}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{A}_{12,\,i,\,j}^{T}$

$\qquad\qquad \boldsymbol{H}_{22,i} \longleftarrow \boldsymbol{H}_{22,i} - \boldsymbol{A}_{12,\,i,\,j}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{A}_{12,\,i,\,j}^{T}$

$\qquad\qquad \boldsymbol{\omega}_5 \longleftarrow \boldsymbol{\omega}_5 - \boldsymbol{A}_{12,ij}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{a}_{2,ij}$; $\boldsymbol{\Omega}_6 \longleftarrow \boldsymbol{\Omega}_6 - \boldsymbol{A}_{12,ij}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{A}_{12,ij}^{T}$

$\qquad \boldsymbol{\omega}_5 \longleftarrow \boldsymbol{\omega}_5 - \boldsymbol{H}_{12,i}\boldsymbol{H}_{22,i}^{-1}\boldsymbol{h}_{2,i}$; $\boldsymbol{\Omega}_6 \longleftarrow \boldsymbol{\Omega}_6 - \boldsymbol{H}_{12,i}\boldsymbol{H}_{22,i}^{-1}\boldsymbol{H}_{12,i}^{T}$

$\boldsymbol{A}^{11} \longleftarrow \boldsymbol{\Omega}_6^{-1}$; $\boldsymbol{x}_1 \longleftarrow \boldsymbol{A}^{11}\boldsymbol{\omega}_5$

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

$\qquad \boldsymbol{x}_{2,i} \longleftarrow \boldsymbol{H}_{22,i}^{-1}(\boldsymbol{h}_{2,i} - \boldsymbol{H}_{12,i}^{T}\boldsymbol{x}_1)$; $\boldsymbol{A}^{12,i} \longleftarrow -(\boldsymbol{H}_{22,i}^{-1}\boldsymbol{H}_{12,i}^{T}\boldsymbol{A}^{11})^{T}$

$\qquad \boldsymbol{A}^{22,i} \longleftarrow \boldsymbol{H}_{22,i}^{-1}(\boldsymbol{I} - \boldsymbol{H}_{12,i}^{T}\boldsymbol{A}^{12,i})$

$\qquad$ For $j = 1, \ldots, n_i$:

$\qquad\qquad \boldsymbol{x}_{2,ij} \longleftarrow \boldsymbol{A}_{22,ij}^{-1}\big(\boldsymbol{a}_{2,ij} - \boldsymbol{A}_{12,ij}^{T}\boldsymbol{x}_1 - \boldsymbol{A}_{12,\,i,\,j}^{T}\boldsymbol{x}_{2,i}\big)$

$\qquad\qquad \boldsymbol{A}^{12,ij} \longleftarrow -\big\{\boldsymbol{A}_{22,ij}^{-1}\big(\boldsymbol{A}_{12,ij}^{T}\boldsymbol{A}^{11} + \boldsymbol{A}_{12,\,i,\,j}^{T}\boldsymbol{A}^{12,i\,T}\big)\big\}^{T}$

$\qquad\qquad \boldsymbol{A}^{12,\,i,\,j} \longleftarrow -\big\{\boldsymbol{A}_{22,ij}^{-1}\big(\boldsymbol{A}_{12,ij}^{T}\boldsymbol{A}^{12,i} + \boldsymbol{A}_{12,\,i,\,j}^{T}\boldsymbol{A}^{22,i}\big)\big\}^{T}$

$\qquad\qquad \boldsymbol{A}^{22,ij} \longleftarrow \boldsymbol{A}_{22,ij}^{-1}\big(\boldsymbol{I} - \boldsymbol{A}_{12,ij}^{T}\boldsymbol{A}^{12,ij} - \boldsymbol{A}_{12,\,i,\,j}^{T}\boldsymbol{A}^{12,\,i,\,j}\big)$

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

$\qquad \big\{\big(\boldsymbol{x}_{2,ij},\, \boldsymbol{A}^{22,ij},\, \boldsymbol{A}^{12,ij},\, \boldsymbol{A}^{12,\,i,\,j},\big) :\ 1 \le i \le m,\ 1 \le j \le n_i\big\}\Big)$

Algorithm 4 SOLVETHREELEVELSPARSELEASTSQUARES *for solving the three-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 (8).*

Input: $\left\{\left(\boldsymbol{b}_{ij}(\tilde{o}_{ij} \times 1),\ \boldsymbol{B}_{ij}(\tilde{o}_{ij} \times p),\ \dot{\boldsymbol{B}}_{ij}(\tilde{o}_{ij} \times q_1),\ \ddot{\boldsymbol{B}}_{ij}(\tilde{o}_{ij} \times q_2)\right) : 1 \le i \le m,\ 1 \le j \le n_i\right\}$

$\boldsymbol{\omega}_7 \longleftarrow$ NULL ; $\boldsymbol{\Omega}_8 \longleftarrow$ NULL

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

$\qquad \boldsymbol{\omega}_9 \longleftarrow$ NULL ; $\boldsymbol{\Omega}_{10} \longleftarrow$ NULL ; $\boldsymbol{\Omega}_{11} \longleftarrow$ NULL

$\qquad$ For $j = 1, \ldots, n_i$:

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

$\qquad\qquad \boldsymbol{d}_{0ij} \longleftarrow \boldsymbol{Q}_{ij}^T\boldsymbol{b}_{ij}$; $\boldsymbol{D}_{0ij} \longleftarrow \boldsymbol{Q}_{ij}^T\boldsymbol{B}_{ij}$; $\dot{\boldsymbol{D}}_{0ij} \longleftarrow \boldsymbol{Q}_{ij}^T\dot{\boldsymbol{B}}_{ij}$

$\qquad\qquad \boldsymbol{d}_{1ij} \longleftarrow$ 1st q_2 rows of $\boldsymbol{d}_{0ij}$; $\boldsymbol{d}_{2ij} \longleftarrow$ remaining rows of $\boldsymbol{d}_{0ij}$; $\boldsymbol{\omega}_9 \longleftarrow \begin{bmatrix} \boldsymbol{\omega}_9 \\ \boldsymbol{d}_{2ij} \end{bmatrix}$

$\qquad\qquad \boldsymbol{D}_{1ij} \longleftarrow$ 1st q_2 rows of $\boldsymbol{D}_{0ij}$; $\boldsymbol{D}_{2ij} \longleftarrow$ remaining rows of $\boldsymbol{D}_{0ij}$; $\boldsymbol{\Omega}_{10} \longleftarrow \begin{bmatrix} \boldsymbol{\Omega}_{10} \\ \boldsymbol{D}_{2ij} \end{bmatrix}$

$\qquad\qquad \dot{\boldsymbol{D}}_{1ij} \longleftarrow$ 1st q_2 rows of $\dot{\boldsymbol{D}}_{0ij}$; $\dot{\boldsymbol{D}}_{2ij} \longleftarrow$ remaining rows of $\dot{\boldsymbol{D}}_{0ij}$; $\boldsymbol{\Omega}_{11} \longleftarrow \begin{bmatrix} \boldsymbol{\Omega}_{11} \\ \dot{\boldsymbol{D}}_{2ij} \end{bmatrix}$

$\qquad$ Decompose $\boldsymbol{\Omega}_{11} = \boldsymbol{Q}_i\begin{bmatrix} \boldsymbol{R}_i \\ \boldsymbol{O} \end{bmatrix}$ such that $\boldsymbol{Q}_i^{-1} = \boldsymbol{Q}_i^T$ and $\boldsymbol{R}_i$ is upper-triangular.

$\qquad \boldsymbol{c}_{0i} \longleftarrow \boldsymbol{Q}_i^T\boldsymbol{\omega}_9$; $\boldsymbol{C}_{0i} \longleftarrow \boldsymbol{Q}_i^T\boldsymbol{\Omega}_{10}$

$\qquad \boldsymbol{c}_{1i} \longleftarrow$ 1st q_1 rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{c}_{2i} \longleftarrow$ remaining rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{\omega}_7 \longleftarrow \begin{bmatrix} \boldsymbol{\omega}_7 \\ \boldsymbol{c}_{2i} \end{bmatrix}$

$\qquad \boldsymbol{C}_{1i} \longleftarrow$ 1st q_1 rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{C}_{2i} \longleftarrow$ remaining rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{\Omega}_8 \longleftarrow \begin{bmatrix} \boldsymbol{\Omega}_8 \\ \boldsymbol{C}_{2i} \end{bmatrix}$

Decompose $\boldsymbol{\Omega}_8 = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{R} \\ \boldsymbol{O} \end{bmatrix}$ so that $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$ and $\boldsymbol{R}$ is upper-triangular.

$\boldsymbol{c} \longleftarrow$ first p rows of $\boldsymbol{Q}^T\boldsymbol{\omega}_7$; $\boldsymbol{x}_1 \longleftarrow \boldsymbol{R}^{-1}\boldsymbol{c}$; $\boldsymbol{A}^{11} \longleftarrow \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$

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

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

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

$\qquad$ For $j = 1, \ldots, n_i$:

$\qquad\qquad \boldsymbol{x}_{2,ij} \longleftarrow \boldsymbol{R}_{ij}^{-1}(\boldsymbol{d}_{1ij} - \boldsymbol{D}_{1ij}\boldsymbol{x}_1 - \dot{\boldsymbol{D}}_{1ij}\boldsymbol{x}_{2,i})$

$\qquad\qquad \boldsymbol{A}^{12,ij} \longleftarrow -\left\{\boldsymbol{R}_{ij}^{-1}(\boldsymbol{D}_{1ij}\boldsymbol{A}^{11} + \dot{\boldsymbol{D}}_{1ij}\boldsymbol{A}^{12,i\,T})\right\}^T$

$\qquad\qquad \boldsymbol{A}^{12,\,i,\,j} \longleftarrow -\left\{\boldsymbol{R}_{ij}^{-1}(\boldsymbol{D}_{1ij}\boldsymbol{A}^{12,i} + \dot{\boldsymbol{D}}_{1ij}\boldsymbol{A}^{22,i})\right\}^T$

$\qquad\qquad \boldsymbol{A}^{22,ij} \longleftarrow \boldsymbol{R}_{ij}^{-1}(\boldsymbol{R}_{ij}^{-T} - \boldsymbol{D}_{1ij}\boldsymbol{A}^{12,ij} - \dot{\boldsymbol{D}}_{1ij}\boldsymbol{A}^{12,\,i,\,j})$

Output: $\left(\boldsymbol{x}_1, \boldsymbol{A}^{11}, \left\{\left(\boldsymbol{x}_{2,i}, \boldsymbol{A}^{22,i}, \boldsymbol{A}^{12,i}\right) : 1 \le i \le m\right\}\right)$
$\qquad\qquad \left\{\left(\boldsymbol{x}_{2,ij}, \boldsymbol{A}^{22,ij}, \boldsymbol{A}^{12,ij}, \boldsymbol{A}^{12,\,i,\,j}\right) : 1 \le i \le m,\ 1 \le j \le n_i\right\}\Big)$

Next, consider the special case where a three-level sparse matrix problem arises as a least squares problem where $\boldsymbol{x}$ is the minimizer of the least squares problem $\|\boldsymbol{b} - \boldsymbol{Bx}\|^2 \equiv$

$(\boldsymbol{b} - \boldsymbol{Bx})^T(\boldsymbol{b} - \boldsymbol{Bx})$ where $\boldsymbol{B}$ is such that $\boldsymbol{A} = \boldsymbol{B}^T\boldsymbol{B}$ has three-level sparse structure. For the special case of $m = 2$, $n_1 = 2$ and $n_2 = 3$ the forms of the $\boldsymbol{B}$ and $\boldsymbol{b}$ matrices are

$$
\boldsymbol{B} \equiv \left[\begin{array}{ccc|ccccc}
\boldsymbol{B}_{11} & \dot{\boldsymbol{B}}_{11} & \ddot{\boldsymbol{B}}_{11} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\
\boldsymbol{B}_{12} & \dot{\boldsymbol{B}}_{12} & \boldsymbol{O} & \ddot{\boldsymbol{B}}_{12} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\
\boldsymbol{B}_{21} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dot{\boldsymbol{B}}_{21} & \ddot{\boldsymbol{B}}_{21} & \boldsymbol{O} & \boldsymbol{O} \\
\boldsymbol{B}_{22} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dot{\boldsymbol{B}}_{22} & \boldsymbol{O} & \ddot{\boldsymbol{B}}_{22} & \boldsymbol{O} \\
\boldsymbol{B}_{23} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dot{\boldsymbol{B}}_{23} & \boldsymbol{O} & \boldsymbol{O} & \ddot{\boldsymbol{B}}_{23}
\end{array}\right] \quad \text{and} \quad \boldsymbol{b} \equiv \left[\begin{array}{c} \boldsymbol{b}_{11} \\ \boldsymbol{b}_{12} \\ \boldsymbol{b}_{21} \\ \boldsymbol{b}_{22} \\ \boldsymbol{b}_{23} \end{array}\right]. \tag{13}
$$

For general $1 \le i \le m$ and $1 \le j \le n_i$, the dimensions of the sub-blocks of $\boldsymbol{b}$ and $\boldsymbol{B}$ are:

$$
\boldsymbol{b}_{ij} \text{ is } \tilde{o}_{ij} \times 1, \quad \boldsymbol{B}_{ij} \text{ is } \tilde{o}_{ij} \times p, \quad \dot{\boldsymbol{B}}_{ij} \text{ is } \tilde{o}_{ij} \times q_1, \quad \text{and} \quad \ddot{\boldsymbol{B}}_{ij} \text{ is } \tilde{o}_{ij} \times q_2. \tag{14}
$$

Here we use $\tilde{o}_{ij}$ rather than o_{ij} to avoid a notational clash with common grouped data dimension notation as used in Section 5. Algorithm 4 provides a QR decomposition-based solution to the three-level sparse matrix least squares problems when the inputs are the matrices listed in (14).

3.3 Additional Matrix Notation

For matrices $\boldsymbol{M}_1, \dots, \boldsymbol{M}_d$ we define:

$$
\operatorname*{stack}_{1 \le i \le d}(\boldsymbol{M}_i) \equiv \left[\begin{array}{c} \boldsymbol{M}_1 \\ \vdots \\ \boldsymbol{M}_d \end{array}\right] \quad \text{and} \quad \operatorname*{blockdiag}_{1 \le i \le d}(\boldsymbol{M}_i) \equiv \left[\begin{array}{cccc} \boldsymbol{M}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{M}_2 & \cdots & \boldsymbol{O} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{M}_d \end{array}\right]
$$

with the first of these definitions requiring that $\boldsymbol{M}_i$, $1 \le i \le d$, each having the same number of columns. Such notation is very useful for defining matrices that appear in higher level random effects models. For example,

$$
\boldsymbol{B} \equiv \left[\operatorname*{stack}_{1 \le i \le m}\left\{\operatorname*{stack}_{1 \le j \le n_i}(\boldsymbol{B}_{ij})\right\} \middle| \operatorname*{blockdiag}_{1 \le i \le m}\left\{\left[\operatorname*{stack}_{1 \le j \le n_i}(\dot{\boldsymbol{B}}_{ij}) \middle| \operatorname*{blockdiag}_{1 \le j \le n_i}(\ddot{\boldsymbol{B}}_{ij})\right]\right\}\right]
$$

$$
\text{and} \quad \boldsymbol{b} \equiv \operatorname*{stack}_{1 \le i \le m}\left\{\operatorname*{stack}_{1 \le j \le n_i}(\boldsymbol{b}_{ij})\right\}. \tag{15}
$$

4 Two-Level Models

We now present streamlined algorithms for two-level linear mixed models. As a stepping stone towards variational inference settings, we first present streamlined best linear unbiased prediction for frequentist linear mixed models.

4.1 Best Linear Unbiased Prediction

Consider the Gaussian response two-level linear mixed model

$$
\boldsymbol{y}_i | \boldsymbol{u}_i \overset{\text{ind.}}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i, \sigma^2\,\boldsymbol{I}), \quad \boldsymbol{u}_i \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad 1 \le i \le m, \tag{16}
$$

where the dimensions of all matrices, for $1 \le i \le m$, are as follows:

$$
\boldsymbol{y}_i \text{ is } n_i \times 1, \quad \boldsymbol{X}_i \text{ is } n_i \times p, \quad \boldsymbol{\beta} \text{ is } p \times 1, \quad \boldsymbol{Z}_i \text{ is } n_i \times q, \quad \boldsymbol{u}_i \text{ is } q \times 1 \text{ and } \boldsymbol{\Sigma} \text{ is } q \times q. \tag{17}
$$

Also, for examples, $\boldsymbol{u}_i \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma})$ is shorthand for the $\boldsymbol{u}_i$ being independently distributed $N(\boldsymbol{0}, \boldsymbol{\Sigma})$ random vectors. Next define the matrices

$$
\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_m \end{bmatrix}, \quad \boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_m \end{bmatrix}, \quad \boldsymbol{Z} \equiv \underset{1 \le i \le m}{\text{blockdiag}}(\boldsymbol{Z}_i), \quad \boldsymbol{u} \equiv \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_m \end{bmatrix}, \quad (18)
$$

$$
\boldsymbol{C} \equiv [\boldsymbol{X} \ \boldsymbol{Z}], \quad \boldsymbol{D}_{\text{BLUP}} \equiv \begin{bmatrix} \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1} \end{bmatrix} \quad \text{and} \quad \boldsymbol{R}_{\text{BLUP}} \equiv \sigma^2 \boldsymbol{I}.
$$

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

$$
\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = (\boldsymbol{C}^T \boldsymbol{R}_{\text{BLUP}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1} \boldsymbol{C}^T \boldsymbol{R}_{\text{BLUP}}^{-1} \boldsymbol{y}
$$
$$
\text{(19)}
$$
$$
\text{and} \quad \text{Cov}\left(\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} - \boldsymbol{u} \end{bmatrix} \right) = (\boldsymbol{C}^T \boldsymbol{R}_{\text{BLUP}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1}.
$$

Despite the simple-looking expressions on the right-hand sides of (19), their storage and computation are major challenges when the sample sizes m and n_i are even moderately large. Fortunately these matrices are sparse and computation of $[\widehat{\boldsymbol{\beta}}^T \ \widehat{\boldsymbol{u}}^T]^T$ can be streamlined. Also, in mixed model applications the entire covariance matrix (19) is usually *not* of interest, but rather only

$$
\text{Cov}(\widehat{\boldsymbol{\beta}}) = \text{top left-hand } p \times p \text{ sub-block of } (\boldsymbol{C}^T \boldsymbol{R}_{\text{BLUP}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1},
$$

$$
\text{Cov}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i) = \text{subsequent } q \times q \text{ diagonal sub-blocks of}
$$
$$
(\boldsymbol{C}^T \boldsymbol{R}_{\text{BLUP}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1} \text{ below Cov}(\widehat{\boldsymbol{\beta}}), \quad 1 \le i \le m, \quad (20)
$$

$$
\text{and } E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)^T\} = \text{subsequent } p \times q \text{ sub-blocks of } (\boldsymbol{C}^T \boldsymbol{R}_{\text{BLUP}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1} \text{ to the}
$$
$$
\text{right of Cov}(\widehat{\boldsymbol{\beta}}), \quad 1 \le i \le m.
$$

Throughout this section we are assuming, without loss of generality, that the variance parameter σ^2 and the covariance matrix $\boldsymbol{\Sigma}$ are known. In practice they are usually unknown and estimated via restricted maximum likelihood (e.g. McCulloch, Searle & Neuhaus, 2008; Section 6.9), in which case the term *empirical* best linear unbiased prediction is used. However, the matrix algebraic challenges that arise in computation of $[\widehat{\boldsymbol{\beta}}^T \ \widehat{\boldsymbol{u}}^T]^T$ and (20) are the same in both cases.

Result 1 facilitates the use of Algorithm 2 for efficient and stable computation of best linear unbiased predictions, and their corresponding covariance matrices, for two-level linear mixed models. It makes use of the matrix square root which, for a symmetric positive definite matrix $\boldsymbol{M}$, is given by $\boldsymbol{M}^{1/2} \equiv \boldsymbol{U} \text{diag}(\sqrt{\boldsymbol{d}}) \boldsymbol{U}^T$ where $\boldsymbol{M} = \boldsymbol{U} \text{diag}(\boldsymbol{d}) \boldsymbol{U}^T$ is the singular value decomposition of $\boldsymbol{M}$.

Result 1. *Computation of $[\widehat{\boldsymbol{\beta}}^T \ \widehat{\boldsymbol{u}}^T]^T$ and each of the sub-blocks of $\text{Cov}([\widehat{\boldsymbol{\beta}}^T \ (\widehat{\boldsymbol{u}} - \boldsymbol{u})^T]^T)$ listed in (20) 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 (9), are

$$
\boldsymbol{b}_i \equiv \begin{bmatrix} \sigma^{-1} \boldsymbol{y}_i \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{B}_i \equiv \begin{bmatrix} \sigma^{-1} \boldsymbol{X}_i \\ \boldsymbol{O} \end{bmatrix} \quad \text{and} \quad \dot{\boldsymbol{B}}_i \equiv \begin{bmatrix} \sigma^{-1} \boldsymbol{Z}_i \\ \boldsymbol{\Sigma}^{-1/2} \end{bmatrix}, \quad 1 \le i \le m,
$$

14

with each of these matrices having $\tilde{n}_i = n_i + q$ rows. The solutions are $\widehat{\boldsymbol{\beta}} = \boldsymbol{x}_1$, $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{11}$ and

$$\widehat{\boldsymbol{u}}_i = \boldsymbol{x}_{2,i}, \;\; \mathrm{Cov}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i) = \boldsymbol{A}^{22,i}, \;\; E\{\boldsymbol{\beta}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)^T\} = \boldsymbol{A}^{12,i}, \quad 1 \le i \le m.$$

Result 1 is derived in Appendix A.1. When combined with Theorem 2 of Nolan & Wand (2019) leads to Algorithm 5 for streamlined best linear unbiased prediction for the two-level Gaussian response model (16).

Section 2.2 of Pinheiro & Bates (2000) makes use of the Result 1 sparse least squares form to achieve streamlined fitting of linear mixed models. However, streamlined computation of covariance matrix components is not covered there.

Algorithm 5 *Streamlined algorithm for obtaining best linear unbiased predictions and corresponding covariance matrix components for the two-level linear mixed model.*

Inputs: $\boldsymbol{y}_i(n_i \times 1)$, $\boldsymbol{X}_i(n_i \times p)$, $\boldsymbol{Z}_i(n_i \times q)$, $1 \le i \le m$; $\quad \sigma^2 > 0$,

$\qquad \boldsymbol{\Sigma}(q \times q)$, symmetric and positive definite.

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

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{y}_i \\ \boldsymbol{0} \end{bmatrix} \;\; ; \;\; \boldsymbol{B}_i \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{X}_i \\ \boldsymbol{0} \end{bmatrix} \;\; ; \;\; \overset{\bullet}{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_i \\ \boldsymbol{\Sigma}^{-1/2} \end{bmatrix}$$

$\mathcal{S}_1 \longleftarrow \textsc{SolveTwoLevelSparseLeastSquares}\Big(\big\{(\boldsymbol{b}_i, \boldsymbol{B}_i, \overset{\bullet}{\boldsymbol{B}}_i) : 1 \le i \le m\big\}\Big)$

$\widehat{\boldsymbol{\beta}} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_1$; $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_1$

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

$\qquad \widehat{\boldsymbol{u}}_i \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_1$; $\mathrm{Cov}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i) \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_1$

$\qquad E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_1$

Outputs: $\widehat{\boldsymbol{\beta}}$, $\mathrm{Cov}(\widehat{\boldsymbol{\beta}})$, $\big\{(\widehat{\boldsymbol{u}}_i, \mathrm{Cov}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i), E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i - \boldsymbol{u}_i)^T\}) : 1 \le i \le m\big\}$

4.2 Mean Field Variational Bayes

Now consider the following Bayesian version of (16):

$$\boldsymbol{y}_i | \boldsymbol{\beta}, \boldsymbol{u}_i, \sigma^2 \overset{\text{ind.}}{\sim} N(\boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}_i, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{u}_i | \boldsymbol{\Sigma} \overset{\text{ind.}}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad 1 \le i \le m,$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \quad \sigma^2 | a_{\sigma^2} \sim \text{Inverse-}\chi^2(\nu_{\sigma^2}, 1/a_{\sigma^2}),$$

$$a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\sigma^2} s_{\sigma^2}^2)), \tag{21}$$

$$\boldsymbol{\Sigma} | \boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}} + 2q - 2, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big),$$

$$\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}), \quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}} \equiv \{\nu_{\boldsymbol{\Sigma}}\text{diag}(s_{\boldsymbol{\Sigma}, 1}^2, \ldots, s_{\boldsymbol{\Sigma}, q}^2)\}^{-1},$$

where the hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p \times 1)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p \times p)$ are such that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is symmetric and positive definite and $\nu_{\sigma^2}, \nu_{\boldsymbol{\Sigma}}, s_{\sigma^2}, s_{\boldsymbol{\Sigma}, 1}, \ldots, s_{\boldsymbol{\Sigma}, q} > 0$. Note that (21) implies that the prior on σ is Half-Cauchy with scale parameter s_{σ^2} and the prior on $\boldsymbol{\Sigma}$ is within the class described in Huang & Wand (2013). As explained in Huang & Wand (2013), such priors allow standard deviation and correlation parameters to have arbitrary non-informativeness. All other matrices in (21) have dimensions and definitions as in (17) and (18).

Now consider the following mean field restriction on the joint posterior density function of all parameters in (21):

$$\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, a_{\sigma^2}, \boldsymbol{A_\Sigma}, \sigma^2, \boldsymbol{\Sigma}|\boldsymbol{y}) \approx \mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, a_{\sigma^2}, \boldsymbol{A_\Sigma})\,\mathfrak{q}(\sigma^2, \boldsymbol{\Sigma}) \qquad (22)$$

where, generically, each $\mathfrak{q}$ represents a density function of the random vector indicated by its argument. Then application of the minimum Kullback-Leibler divergence equations (e.g. equation (10.9) of Bishop, 2006) leads to the optimal $\mathfrak{q}$-density functions for the parameters of interest being as follows:

$$\mathfrak{q}^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ has a } N\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}\big) \text{ density function,}$$
$$\mathfrak{q}^*(\sigma^2) \text{ has an Inverse-}\chi^2\big(\xi_{\mathfrak{q}(\sigma^2)}, \lambda_{\mathfrak{q}(\sigma^2)}\big) \text{ density function} \qquad (23)$$
$$\text{and } \mathfrak{q}^*(\boldsymbol{\Sigma}) \text{ has an Inverse-G-Wishart}(G_{\text{full}}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}) \text{ density function.}$$

The optimal $\mathfrak{q}$-density parameters are determined via an iterative coordinate ascent algorithm, with details deferred to Appendix A.4. Algorithm 2 of Lee & Wand (2016) a is naïve mean field variational Bayes algorithm for a class of two-level Gaussian response linear mixed models that includes model (21) as a special case. Subsequent algorithms in Lee & Wand (2016) achieve streamlining. In the current article, we offer an alternative approach, based on Algorithms 6 and 11, that handle higher level random effects in a natural way.

Note that updates for $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ may be written

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T \boldsymbol{R}_{\text{MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1}(\boldsymbol{C}^T \boldsymbol{R}_{\text{MFVB}}^{-1} \boldsymbol{y} + \boldsymbol{o}_{\text{MFVB}})$$
$$\text{and} \quad \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T \boldsymbol{R}_{\text{MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1} \qquad (24)$$

where

$$\boldsymbol{R}_{\text{MFVB}} \equiv \mu_{\mathfrak{q}(1/\sigma^2)}^{-1}\boldsymbol{I}, \quad \boldsymbol{D}_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_m \otimes \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} \end{bmatrix} \quad \text{and} \quad \boldsymbol{o}_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}. \quad (25)$$

For increasingly large sample sizes the matrix $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ becomes untenably massive. Fortunately, only the following relatively small sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ are required for variational inference concerning σ^2 and $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} = \text{top left-hand } p \times p \text{ sub-block of } (\boldsymbol{C}^T \boldsymbol{R}_{\text{MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1},$$
$$\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} = \text{subsequent } q \times q \text{ diagonal sub-blocks of } (\boldsymbol{C}^T \boldsymbol{R}_{\text{MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1}$$
$$\text{below } \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}, 1 \le i \le m, \text{ and} \qquad (26)$$
$$E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\beta)})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\} = \text{subsequent } p \times q \text{ sub-blocks of}$$
$$(\boldsymbol{C}^T \boldsymbol{R}_{\text{MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1} \text{ to the right of } \boldsymbol{\Sigma}_{\mathfrak{q}(\beta)}, 1 \le i \le m.$$

For a streamlined mean field variational Bayes algorithm, we appeal to:

Result 2. *The mean field variational Bayes updates of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and each of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ in (26) is expressible as a two-level sparse matrix least squares problem of the form:*

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

where $\boldsymbol{b}$ and the non-zero sub-blocks of $\boldsymbol{B}$, according to the notation in (9), are, for $1 \le i \le m$,

$$\boldsymbol{b}_i \equiv \begin{bmatrix} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{y}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{B}_i \equiv \begin{bmatrix} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{X}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \end{bmatrix} \quad \text{and} \quad \dot{\boldsymbol{B}}_i \equiv \begin{bmatrix} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{Z}_i \\ \boldsymbol{O} \\ \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix},$$

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

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

and

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

Result 2 implies that the SOLVETWOLEVELSPARSELEASTSQUARES algorithm listed in Algorithm 2 applies for handling the $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ sub-block updates. A derivation is in Appendix A.3. This results in Algorithm 6 for streamlined mean field variational Bayes for the two-level Gaussian response model. Its derivation is given in Appendix A.4.

Algorithm 6 uses the mean field variational Bayes approximate marginal log-likelihood $\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\}$ in its stopping criterion. For model (21) this is given by

$$\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\} = E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma^2,a_{\sigma^2},\boldsymbol{\Sigma},\boldsymbol{A}_{\boldsymbol{\Sigma}}) - \mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,a_{\sigma^2},\boldsymbol{\Sigma},\boldsymbol{A}_{\boldsymbol{\Sigma}})\}. \qquad (27)$$

An explicit streamlined expression for $\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\}$ and corresponding derivation is given in Appendix A.5.

4.3 Variational Message Passing

We now turn attention to the variational message passing alternative. Note that the joint density function of all of the random variables and random vectors in the Bayesian two-level Gaussian response linear mixed model (21) admits the following factorization:

$$\mathfrak{p}(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\boldsymbol{\Sigma},a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}}) = \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\mathfrak{p}(\sigma^2|a_{\sigma^2})\mathfrak{p}(a_{\sigma^2})\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\mathfrak{p}(\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}})\mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}}). \qquad (28)$$

Figure 4 shows a factor graph representation of (28) with color-coding of *fragment* types, according to the nomenclature in Wand (2017).
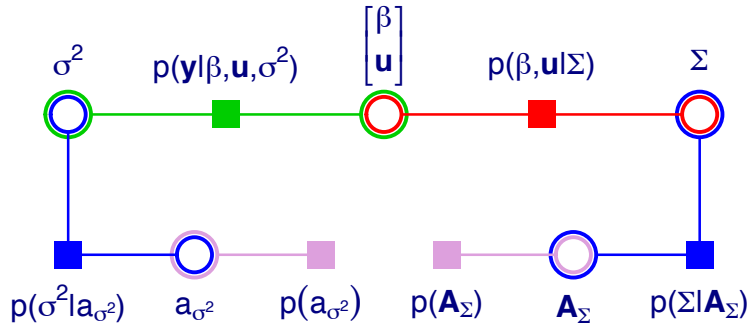

Figure 4: *Factor graph representation of the Bayesian two-level Gaussian response linear mixed model (21).*

Each of these fragments is treated in Section 4.1 of Wand (2017). However, the updates for the Gaussian likelihood fragment, shown in green in Figure 4, and the Gaussian penalization fragment, shown in blue in Figure 4, are given in simple naïve forms in Wand (2017) without matrix algebraic streamlining. The next two subsections overcome this deficiency.

4.4 Streamlined Gaussian Likelihood Fragment Updates

We now focus on the Gaussian likelihood fragment, shown in green in Figure 4. As presented in Section 4.1.5 of Wand (2017), the messages passed between $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)$ and

Algorithm 6 *QR-decomposition-based streamlined algorithm for obtaining mean field variational Bayes approximate posterior density functions for the parameters in the two-level linear mixed model (21) with product density restriction (22).*

Data Inputs: $\boldsymbol{y}_i(n_i \times 1)$, $\boldsymbol{X}_i(n_i \times p)$, $\boldsymbol{Z}_i(n_i \times q)$, $1 \le i \le m$.

Hyperparameter Inputs: $\boldsymbol{\mu_\beta}(p \times 1)$, $\boldsymbol{\Sigma_\beta}(p \times p)$ symmetric and positive definite,
$$s_{\sigma^2}, \nu_{\sigma^2}, s_{\boldsymbol{\Sigma}, 1}, \ldots, s_{\boldsymbol{\Sigma}, q}, \nu_{\boldsymbol{\Sigma}} > 0.$$

Initialize: $\mu_{\mathsf{q}(1/\sigma^2)} > 0$, $\mu_{\mathsf{q}(1/a_{\sigma^2})} > 0$, $M_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})}(q \times q)$, $M_{\mathsf{q}(A_{\boldsymbol{\Sigma}}^{-1})}(q \times q)$ both symmetric and positive definite.

$\xi_{\mathsf{q}(\sigma^2)} \longleftarrow \nu_{\sigma^2} + \sum_{i=1}^m n_i$; $\xi_{\mathsf{q}(\boldsymbol{\Sigma})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + 2q - 2 + m$; $\xi_{\mathsf{q}(a_{\sigma^2})} \longleftarrow \nu_{\sigma^2} + 1$; $\xi_{\mathsf{q}(A_{\boldsymbol{\Sigma}})} \longleftarrow \nu_{\boldsymbol{\Sigma}} + q$

Cycle:

For $i = 1, \ldots, m$:
$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2}\boldsymbol{y}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu_\beta} \\ \boldsymbol{0} \end{bmatrix}, \ \boldsymbol{B}_i \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2}\boldsymbol{X}_i \\ m^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{0} \end{bmatrix}, \ \dot{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2}\boldsymbol{Z}_i \\ \boldsymbol{0} \\ M_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix}.$$

$\mathcal{S}_2 \longleftarrow$ SOLVETWOLEVELSPARSELEASTSQUARES$\Big(\big\{ (\boldsymbol{b}_i, \boldsymbol{B}_i, \dot{\boldsymbol{B}}_i) : 1 \le i \le m \big\} \Big)$

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_2$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_2$

$\lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \mu_{\mathsf{q}(1/a_{\sigma^2})}$; $\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma})} \longleftarrow M_{\mathsf{q}(A_{\boldsymbol{\Sigma}}^{-1})}$

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

$\quad \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_2$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_2$

$\quad E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_2$

$\quad \lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \lambda_{\mathsf{q}(\sigma^2)} + \big\| \boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_i \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} \big\|^2$

$\quad \lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \lambda_{\mathsf{q}(\sigma^2)} + \mathrm{tr}(\boldsymbol{X}_i^T \boldsymbol{X}_i \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}) + \mathrm{tr}(\boldsymbol{Z}_i^T \boldsymbol{Z}_i \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)})$

$\quad \lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \lambda_{\mathsf{q}(\sigma^2)} + 2\,\mathrm{tr}\big[\boldsymbol{Z}_i^T \boldsymbol{X}_i E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} \big]$

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

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

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

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

$M_{\mathsf{q}(A_{\boldsymbol{\Sigma}}^{-1})} \longleftarrow \xi_{\mathsf{q}(A_{\boldsymbol{\Sigma}})}\boldsymbol{\Lambda}_{\mathsf{q}(A_{\boldsymbol{\Sigma}})}^{-1}.$

until the increase in $\log\{\underline{\mathfrak{p}}(\boldsymbol{y}; \mathsf{q})\}$ is negligible.

Outputs: $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}, \big\{\big(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}, E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\}\big) : 1 \le i \le m\big\}$
$\quad \xi_{\mathsf{q}(\sigma^2)}, \lambda_{\mathsf{q}(\sigma^2)}, \xi_{\mathsf{q}(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma})}$

$(\boldsymbol{\beta}, \boldsymbol{u})$ involve Multivariate Normal distributions with natural parameter vectors containing

$$p + mq + \tfrac{1}{2}(p + mq)(p + mq + 1) \tag{29}$$

unique entries. Since the sizes of these vectors grow quadratically with the number of groups, message passing suffers from burdensome storage and computational demands. We overcome this problem by noticing that messages passed to and from $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)$ are within *reduced* Multivariate Normal families.

Note that the full conditional density function of $(\boldsymbol{\beta}, \boldsymbol{u})$ is Multivariate Normal with

inverse covariance matrix

$$\text{Cov}(\boldsymbol{\beta}, \boldsymbol{u}|\text{rest})^{-1} = \sigma^{-2}\boldsymbol{C}^T\boldsymbol{C} + \text{blockdiag}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}, \boldsymbol{I}_m \otimes \boldsymbol{\Sigma}^{-1}),$$

where 'rest' denotes all other random variables in the model, is a two-level sparse matrix. The same is true for $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}^{-1}$, the inverse covariance matrix of the mean field approximate posterior density function of $(\boldsymbol{\beta}, \boldsymbol{u})$. In the variational message passing approach this sparseness transfers to reduced exponential family forms being sufficient. For example, in the case of $p = q = 2$ the messages passed between $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)$ and $(\boldsymbol{\beta}, \boldsymbol{u}) = (\beta_0, \beta_1, u_{10}, u_{11}, \ldots, u_{m0}, u_{m1})$ have the generic exponential family forms:

$$\exp\left\{\eta_{\beta_0} + \eta_{\beta_1}\beta_1 + \sum_{i=1}^m (\eta_{u_{i0}} u_{i0} + \eta_{u_{i1}} u_{i1}) + \eta_{\beta_0^2}\beta_0^2 + \eta_{\beta_1^2}\beta_1^2 + \sum_{i=1}^m (\eta_{u_{i0}^2}\, u_{i0}^2 + \eta_{u_{i1}^2}\, u_{i1}^2) \right.$$
$$\left. + \sum_{i=1}^m (\eta_{\beta_0 u_{i0}}\beta_0\, u_{i0} + \eta_{\beta_0 u_{i1}}\beta_0\, u_{i1} + \eta_{\beta_1 u_{i0}}\beta_1\, u_{i0} + \eta_{\beta_1 u_{i1}}\beta_1\, u_{i1}) \right\}. \tag{30}$$

Therefore, it is natural to insist that all messages passed to $(\boldsymbol{\beta}, \boldsymbol{u})$ from factors outside of the two-level Gaussian likelihood fragment are within the same reduced exponential family. Under such a conjugacy constraint, the natural parameter vectors of messages passed to and from $(\boldsymbol{\beta}, \boldsymbol{u})$ have length

$$p + \tfrac{1}{2}\,p(p+1) + m\{q + \tfrac{1}{2}\,q(q+1) + pq\}$$

which is linear in m and considerably lower than (29) when the number of groups is large. The reduced exponential family has as an attractive graph theoretic representation. The full Multivariate Normal distribution, in which sparseness is ignored, has dimension $p + mq$. The probabilistic undirected graph that respects independence of any pair of random variables conditional on the rest for the $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution is an undirected graph with an edge between the ℓth and ℓ'th nodes if and only if $(\boldsymbol{\Sigma}^{-1})_{\ell\ell'} \neq 0$ (e.g. Rue & Held, 2005). The restricted exponential family corresponds to removal of edges in a fully connected $(p + mq)$-node graph. Figure 5 depicts the reduced graph in the case of $p = q = 2$ and $m = 4$. The fully connected graph has 45 edges, whereas the reduced graph corresponding to the restricted exponential family has only 21 edges. For general p, q and m the numbers of edges are, respectively, $\tfrac{1}{2}(p + mq)(p + mq - 1)$ and $\tfrac{1}{2}\,p(p-1) + m\{\tfrac{1}{2}\,q(q-1) + pq\}$. So, for example, if $p = q = 2$ and $m = 10,000$ then the number of edges in the reduced graph is about 50,000 compared with about 200 million in the full graph.

The message from $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)$ to $(\boldsymbol{\beta}, \boldsymbol{u})$ is

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to (\boldsymbol{\beta}, \boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{u}) =$$

$$\exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1 \le i \le m}{\text{stack}} \begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to (\boldsymbol{\beta}, \boldsymbol{u})} \right\} \tag{31}$$

with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to (\boldsymbol{\beta}, \boldsymbol{u})}$ of length

$$p + \tfrac{1}{2}p(p+1) + m\{q + \tfrac{1}{2}q(q+1) + pq\}. \tag{32}$$

Under conjugacy, the reverse message $m_{(\boldsymbol{\beta}, \boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)}(\boldsymbol{\beta}, \boldsymbol{u})$ has the same algebraic form as (31) with natural parameter vector $\boldsymbol{\eta}_{(\boldsymbol{\beta}, \boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)}$ also of length (32).
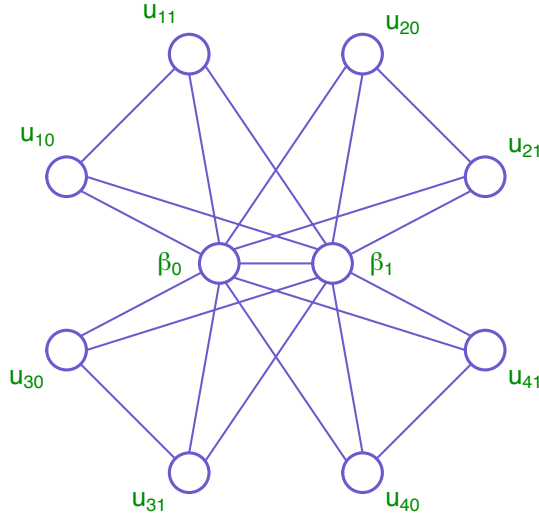
Figure 5: *Undirected probabilistic graph with edges coding the conditional dependencies of the entries of $(\boldsymbol{\beta}, \boldsymbol{u})$ given the rest for the case $p = q = 2$ and $m = 4$.*

Result 3. *The variational message passing updates of the quantities $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}$, $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}$, $1 \le i \le m$, and the sub-blocks of $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ listed in (26) with q-density expectations with respect to the normalization of*

$$m_{\mathfrak{p}(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to (\boldsymbol{\beta}, \boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{u}) \, m_{(\boldsymbol{\beta}, \boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)}(\boldsymbol{\beta}, \boldsymbol{u})$$

are expressible as a two-level sparse matrix problem with

$$\boldsymbol{A} = -2 \begin{bmatrix} \operatorname{vec}^{-1}(\boldsymbol{D}_p^{+T} \boldsymbol{\eta}_{1,2}) & \left[\frac{1}{2} \underset{1 \le i \le m}{\operatorname{stack}} \{\operatorname{vec}^{-1}(\boldsymbol{\eta}_{2,3,i})^T\} \right]^T \\ \frac{1}{2} \underset{1 \le i \le m}{\operatorname{stack}} \{\operatorname{vec}^{-1}(\boldsymbol{\eta}_{2,3,i})^T\} & \underset{1 \le i \le m}{\operatorname{blockdiag}} \{\operatorname{vec}^{-1}(\boldsymbol{D}_q^{+T} \boldsymbol{\eta}_{2,2,i})\} \end{bmatrix}$$

and

$$\boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{\eta}_{1,1} \\ \underset{1 \le i \le m}{\operatorname{stack}} (\boldsymbol{\eta}_{2,1,i}) \end{bmatrix} \quad where \quad \begin{bmatrix} \boldsymbol{\eta}_{1,1} \ (p \times 1) \\ \boldsymbol{\eta}_{1,2} \ (\frac{1}{2}p(p+1) \times 1) \\ \underset{1 \le i \le m}{\operatorname{stack}} \begin{bmatrix} \boldsymbol{\eta}_{2,1,i} & (q \times 1) \\ \boldsymbol{\eta}_{2,2,i} & (\frac{1}{2}q(q+1) \times 1) \\ \boldsymbol{\eta}_{2,3,i} & (pq \times 1) \end{bmatrix} \end{bmatrix}$$

is the partitioning of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \leftrightarrow (\boldsymbol{\beta}, \boldsymbol{u})}$ that defines $\boldsymbol{\eta}_{1,1}$, $\boldsymbol{\eta}_{1,2}$ and $\{(\boldsymbol{\eta}_{2,1,i}, \boldsymbol{\eta}_{2,2,i}, \boldsymbol{\eta}_{2,3,i}) : 1 \le i \le m\}$. The solutions are $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} = \boldsymbol{x}_1$, $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$ and

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

Remark. Variational message passing differs from mean field variational Bayes in that its two-level sparse matrix problem is not expressible in a least squares form.

The process of converting a generic reduced natural parameter vector $\boldsymbol{\eta}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ to the corresponding $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ vector and important sub-blocks of $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})}$, as illustrated by Result 3, is fundamental to streamlining of variational message passing for two-level linear mixed models. We call this procedure the TwoLevelNaturalToCommonParameters algorithm and list required steps as Algorithm 7.

Algorithm 7 *The* TwoLevelNaturalToCommonParameters *algorithm for conversion of a two-level reduced natural parameter vector to its corresponding common parameters.*

Inputs: $p, q, m, \boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$

$\boldsymbol{\omega}_{12} \longleftarrow$ first p entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$

$\boldsymbol{\omega}_{13} \longleftarrow$ next $\frac{1}{2}p(p+1)$ entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$; $\boldsymbol{\Omega}_{14} \longleftarrow -2\text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\omega}_{13})$

$i_{\text{stt}} \longleftarrow p + \frac{1}{2}p(p+1) + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q - 1$;

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

    $\boldsymbol{\omega}_{15i} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

    $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + \frac{1}{2}q(q+1) - 1$

    $\boldsymbol{\omega}_{16} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

    $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq - 1$

    $\boldsymbol{\omega}_{17} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

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

    $\boldsymbol{\Omega}_{18i} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_q^{+T}\boldsymbol{\omega}_{16})$; $\boldsymbol{\Omega}_{19i} \longleftarrow -\text{vec}_{p \times q}^{-1}(\boldsymbol{\omega}_{17})$

$\mathcal{S}_3 \longleftarrow$ SolveTwoLevelSparseMatrix$\Big(\boldsymbol{\omega}_{12}, \boldsymbol{\Omega}_{14}, \big\{(\boldsymbol{\omega}_{15i}, \boldsymbol{\Omega}_{18i}, \boldsymbol{\Omega}_{19i}) : 1 \leq i \leq m\big\}\Big)$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_3$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_3$

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

    $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_3$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_3$

    $E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_3$

Outputs: $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}, \big\{\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}, E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}\big) : 1 \leq i \leq m\big\}$

It is easily shown (Appendix A.8) that messages between $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)$ and σ^2 have Inverse Chi-Squared forms. For example,

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to \sigma^2}(\sigma^2) = \exp\left\{ \begin{bmatrix} 1/\sigma^2 \\ \log(\sigma^2) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to \sigma^2} \right\}. \tag{33}$$

Algorithm 8 lists parameter updates for the two-level Gaussian likelihood fragment with streamlining according to the restricted exponential family form (31). Note that it makes use of SolveTwoLevelSparseMatrix (Algorithm 1) since the natural parameter updates correspond to a two-level sparse matrix problem *without* least squares representation. Appendix A.8 provides details on the derivation of Algorithm 8.

As in Wand (2017), Algorithm 8 uses the notation

$$\boldsymbol{\eta}_{f \leftrightarrow \theta} \equiv \boldsymbol{\eta}_{f \to \theta} + \boldsymbol{\eta}_{\theta \to f}. \tag{34}$$

4.5 Streamlined Gaussian Penalization Fragment Updates

Next we turn our attention to the Gaussian penalization fragment when the random effects vector has two-level structure. The relevant fragment is shown in red in Figure 4.

As shown in Appendix A.10, the message from $\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})$ to $(\boldsymbol{\beta}, \boldsymbol{u})$ has the generic form (30) but with even more vanishing terms than the message passed from $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)$. However, with conjugacy in mind, we work with messages having the same form as (31).

Algorithm 8 *The inputs, updates and outputs of the matrix algebraic streamlined Gaussian likelihood fragment for two-level models.*

Data Inputs: $\boldsymbol{y}_i(n_i \times 1)$, $\boldsymbol{X}_i(n_i \times p)$, $\boldsymbol{Z}_i(n_i \times q)$, $1 \le i \le m$

Parameter Inputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to \sigma^2}$,
$\qquad \boldsymbol{\eta}_{\sigma^2 \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$

Updates:

$$\mu_{\mathsf{q}(1/\sigma^2)} \longleftarrow \left(\left(\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2}\right)_1 + 1 \right) \Big/ \left(\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2}\right)_2$$

$$\mathcal{S}_4 \longleftarrow \text{TwoLevelNaturalToCommonParameters}\left(p, q, m, \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}\right)$$

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}$ component of $\mathcal{S}_4$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}$ component of $\mathcal{S}_4$

$\boldsymbol{\omega}_{20} \longleftarrow \boldsymbol{0}_p$; $\boldsymbol{\omega}_{21} \longleftarrow \boldsymbol{0}_{\frac{1}{2} p(p+1)}$; $\boldsymbol{\omega}_{22} \longleftarrow 0$

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

$\qquad \boldsymbol{\omega}_{20} \longleftarrow \boldsymbol{\omega}_{20} + \boldsymbol{X}_i^T \boldsymbol{y}_i$; $\boldsymbol{\omega}_{21} \longleftarrow \boldsymbol{\omega}_{21} - \frac{1}{2}\boldsymbol{D}_p^T \text{vec}(\boldsymbol{X}_i^T \boldsymbol{X}_i)$

$\qquad \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}$ component of $\mathcal{S}_4$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}$ component of $\mathcal{S}_4$

$\qquad E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} \longleftarrow E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\}$ component

$\qquad\qquad\qquad\qquad$ of $\mathcal{S}_4$

$\qquad \boldsymbol{\omega}_{22} \longleftarrow \boldsymbol{\omega}_{22} - \frac{1}{2}\|\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_i \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}\|^2$

$\qquad \boldsymbol{\omega}_{22} \longleftarrow \boldsymbol{\omega}_{22} - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}\boldsymbol{X}_i^T \boldsymbol{X}_i) - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}\boldsymbol{Z}_i^T \boldsymbol{Z}_i)$

$\qquad\qquad -\text{tr}[\{\boldsymbol{Z}_i^T \boldsymbol{X}_i E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\}]$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \mu_{\mathsf{q}(1/\sigma^2)} \begin{bmatrix} \boldsymbol{\omega}_{20} \\ \boldsymbol{\omega}_{21} \\ \underset{1 \le i \le m}{\text{stack}} \begin{bmatrix} \boldsymbol{Z}_i^T \boldsymbol{y}_i \\ -\frac{1}{2}\boldsymbol{D}_q^T \text{vec}(\boldsymbol{Z}_i^T \boldsymbol{Z}_i) \\ -\text{vec}(\boldsymbol{X}_i^T \boldsymbol{Z}_i) \end{bmatrix} \end{bmatrix}$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to \sigma^2} \longleftarrow \begin{bmatrix} -\frac{1}{2}\sum_{i=1}^m n_i \\ \boldsymbol{\omega}_{22} \end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to \sigma^2}$.

This implies that

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) = \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1 \le i \le m}{\text{stack}} \begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}) \to (\boldsymbol{\beta},\boldsymbol{u})} \right\}$$

with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}$ also of length (32). The reverse message has an analogous form.

Result 4. *The variational message passing updates of the quantities $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}$, $1 \le i \le m$,*

with q-density expectations with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \rightarrow (\boldsymbol{\beta}, \boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{u}) \, m_{(\boldsymbol{\beta}, \boldsymbol{u}) \rightarrow \mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{\beta}, \boldsymbol{u})$$

are expressible as a two-level sparse matrix problem with

$$\boldsymbol{A} = -2 \begin{bmatrix} \mathrm{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\eta}_{1,2}) & \left[\frac{1}{2}\underset{1 \leq i \leq m}{\mathrm{stack}}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{2,3,i})^T\}\right]^T \\ \frac{1}{2}\underset{1 \leq i \leq m}{\mathrm{stack}}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{2,3,i})^T\} & \underset{1 \leq i \leq m}{\mathrm{blockdiag}}\{\mathrm{vec}^{-1}(\boldsymbol{D}_q^{+T}\boldsymbol{\eta}_{2,2,i})\} \end{bmatrix}$$

and

$$\boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{\eta}_{1,1} \\ \underset{1 \leq i \leq m}{\mathrm{stack}}(\boldsymbol{\eta}_{2,1,i}) \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} \boldsymbol{\eta}_{1,1} \ (p \times 1) \\ \boldsymbol{\eta}_{1,2} \ (\frac{1}{2}p(p+1) \times 1) \\ \underset{1 \leq i \leq m}{\mathrm{stack}} \begin{bmatrix} \boldsymbol{\eta}_{2,1,i} & (q \times 1) \\ \boldsymbol{\eta}_{2,2,i} & (\frac{1}{2}q(q+1) \times 1) \\ \boldsymbol{\eta}_{2,3,i} & (pq \times 1) \end{bmatrix} \end{bmatrix}$$

is the partitioning of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \leftrightarrow (\boldsymbol{\beta}, \boldsymbol{u})}$ that defines $\boldsymbol{\eta}_{1,1}$, $\boldsymbol{\eta}_{1,2}$ and $\{(\boldsymbol{\eta}_{2,1,i}, \boldsymbol{\eta}_{2,2,i}, \boldsymbol{\eta}_{2,3,i}) : 1 \leq i \leq m\}$. The solutions are

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)} = \boldsymbol{x}_{2,i} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} = \boldsymbol{A}^{22,i}, \quad 1 \leq i \leq m.$$

As shown in Appendix A.8, the message from $\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})$ to $\boldsymbol{\Sigma}$ has the Inverse-G-Wishart form

$$m_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \rightarrow \boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) = \exp\left\{ \begin{bmatrix} \log|\boldsymbol{\Sigma}| \\ \mathrm{vech}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \rightarrow \boldsymbol{\Sigma}} \right\}.$$

Conjugacy considerations dictate that the message from $\boldsymbol{\Sigma}$ to $\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})$ is within the same exponential family.

Algorithm 9 lists the natural parameter updates for the Gaussian penalization fragment for two-level random effects. Notation such as $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \leftrightarrow \boldsymbol{\Sigma}}$ is as defined by (34). See Appendix A.8 for its derivation.

4.6 q-Density Determination After Variational Message Passing Convergence

After convergence of the variational message passing iterations, determination of q-density parameters of interest requires some additional non-trivial steps, essentially involving mapping particular natural parameter vectors to common parameters of interest. We will explain this in the context of inference for the parameters in (21) and its Figure 4 factor graph representation.

For the fixed and random effects parameters we need to first carry out:

$$\boldsymbol{\eta}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})} \longleftarrow \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \rightarrow (\boldsymbol{\beta}, \boldsymbol{u})} + \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \rightarrow (\boldsymbol{\beta}, \boldsymbol{u})}$$

$$\mathcal{S}_6 \longleftarrow \textsc{TwoLevelNaturalToCommonParameters}\left(p, q, m, \boldsymbol{\eta}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})}\right)$$

and then unpack $\mathcal{S}_6$ to obtain the mean and important covariance matrix sub-blocks:

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}, \left\{\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}, E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} : 1 \leq i \leq m\right\}.$$

of the $N(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta}, \boldsymbol{u})})$ optimal q-density function.

Algorithm 9 *The inputs, updates and outputs of the matrix algebraic streamlined Gaussian penalization fragment for two-level random effects.*

Hyperparameter Inputs: $\boldsymbol{\mu_\beta}(p \times 1)$, $\boldsymbol{\Sigma_\beta}(p \times p)$, m, q

Parameter Inputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to (\boldsymbol{\beta}, \boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta}, \boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}$, $\boldsymbol{\eta}_{\boldsymbol{\Sigma} \to \mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})}$

Updates:

$\omega_{23} \longleftarrow$ first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \leftrightarrow \boldsymbol{\Sigma}}$; $\omega_{24} \longleftarrow$ remaining entries of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \leftrightarrow \boldsymbol{\Sigma}}$

$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} \longleftarrow \left\{\omega_{23} + \tfrac{1}{2}(q+1)\right\}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_q^{+T}\boldsymbol{\omega}_{24}\right)\right\}^{-1}$

$\mathcal{S}_5 \longleftarrow \text{TwoLevelNaturalToCommonParameters}\left(p, q, m, \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \leftrightarrow (\boldsymbol{\beta}, \boldsymbol{u})}\right)$

$\boldsymbol{\omega}_{25} \longleftarrow \boldsymbol{0}_{\frac{1}{2}q(q+1)}$

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}$ component of $\mathcal{S}_5$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}$ component of $\mathcal{S}_5$

$\boldsymbol{\omega}_{25} \longleftarrow \boldsymbol{\omega}_{25} - \tfrac{1}{2}\boldsymbol{D}_q^T\text{vec}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to (\boldsymbol{\beta}, \boldsymbol{u})} \longleftarrow \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\ \underset{1 \leq i \leq m}{\text{stack}}\begin{bmatrix} \boldsymbol{0}_q \\ -\tfrac{1}{2}\boldsymbol{D}_q^T\text{vec}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\right) \\ \boldsymbol{0}_{pq} \end{bmatrix} \end{bmatrix}$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}} \longleftarrow \begin{bmatrix} -\tfrac{1}{2}m \\ \boldsymbol{\omega}_{25} \end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to (\boldsymbol{\beta}, \boldsymbol{u})}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}}$.

The error variance σ^2 has its optimal $\mathfrak{q}$-density function being that of an Inverse-$\chi^2\left(\xi_{\mathfrak{q}(\sigma^2)}, \lambda_{\mathfrak{q}(\sigma^2)}\right)$ distribution, and its parameters are determined from the steps:

$$\boldsymbol{\eta}_{\mathfrak{q}(\sigma^2)} \longleftarrow \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) \to \sigma^2} + \boldsymbol{\eta}_{\mathfrak{p}(\sigma^2|a_{\sigma^2}) \to \sigma^2}$$

$$\xi_{\mathfrak{q}(\sigma^2)} \longleftarrow -2\left(\boldsymbol{\eta}_{\mathfrak{q}(\sigma^2)}\right)_1 - 2, \;\; \lambda_{\mathfrak{q}(\sigma^2)} \longleftarrow -2\left(\boldsymbol{\eta}_{\mathfrak{q}(\sigma^2)}\right)_2$$

where $\left(\boldsymbol{\eta}_{\mathfrak{q}(\sigma^2)}\right)_j$ denotes the jth entry of the vector $\boldsymbol{\eta}_{\mathfrak{q}(\sigma^2)}$ for $j = 1, 2$.

Finally, the random effects covariance matrix $\boldsymbol{\Sigma}$ has its optimal $\mathfrak{q}$-density function being that of an Inverse-G-Wishart$\left(G_{\text{full}}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\right)$ distribution. The steps for determining its parameters after variational message passing convergence are:

$$\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}) \to \boldsymbol{\Sigma}} + \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\Sigma}|A_{\boldsymbol{\Sigma}}) \to \boldsymbol{\Sigma}}$$

$$\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow -2\left(\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}\right)_1 - 2, \;\; \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow -2\,\text{vec}^{-1}\left(\boldsymbol{D}_q^{+T}\left(\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}\right)_2\right)$$

where $\left(\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}\right)_1$ denotes the first entry of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}$ and $\left(\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}\right)_2$ denotes its remaining entries.

5 Three-Level Models

We now return to the three-level situation illustrated by Figure 1 and derived algorithms for streamlined variational inference based on Algorithms 3 and 4. As a prelude to the

challenging mean field variational Bayes and variational message passing problems, we cover the best linear unbiased prediction problem next.

5.1 Best Linear Unbiased Prediction

The Gaussian response three-level linear mixed model is

$$\boldsymbol{y}_{ij}|\boldsymbol{u}_i^{\text{L1}}, \boldsymbol{u}_{ij}^{\text{L2}} \stackrel{\text{ind.}}{\sim} N(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{u}_i^{\text{L1}} + \boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{u}_{ij}^{\text{L2}}, \sigma^2\,\boldsymbol{I}),$$

$$\left[\begin{array}{c}\boldsymbol{u}_i^{\text{L1}}\\\boldsymbol{u}_{ij}^{\text{L2}}\end{array}\right] \stackrel{\text{ind.}}{\sim} N\left(\left[\begin{array}{c}\boldsymbol{0}\\\boldsymbol{0}\end{array}\right], \left[\begin{array}{cc}\boldsymbol{\Sigma}^{\text{L1}} & \boldsymbol{O}\\\boldsymbol{O} & \boldsymbol{\Sigma}^{\text{L2}}\end{array}\right]\right), \quad 1 \le i \le m,\ 1 \le j \le n_i$$

where the dimensions of all matrices, for $1 \le i \le m, 1 \le j \le n_i$, are as follows:

$$\boldsymbol{y}_{ij} \text{ is } o_{ij} \times 1, \quad \boldsymbol{X}_{ij} \text{ is } o_{ij} \times p, \quad \boldsymbol{\beta} \text{ is } p \times 1, \quad \boldsymbol{Z}_{ij}^{\text{L1}} \text{ is } o_{ij} \times q_1, \quad \boldsymbol{u}_i^{\text{L1}} \text{ is } q_1 \times 1$$

$$\boldsymbol{Z}_{ij}^{\text{L2}} \text{ is } o_{ij} \times q_2, \quad \boldsymbol{u}_{ij}^{\text{L2}} \text{ is } q_2 \times 1, \quad \boldsymbol{\Sigma}^{\text{L1}} \text{ is } q_1 \times q_1 \text{ and } \boldsymbol{\Sigma}^{\text{L2}} \text{ is } q_2 \times q_2.$$

Next let the sub-matrices $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$, $1 \le i \le m$, be defined by

$$\boldsymbol{y}_i \equiv \left[\begin{array}{c}\boldsymbol{y}_{i1}\\\vdots\\\boldsymbol{y}_{in_i}\end{array}\right] \quad \text{and} \quad \boldsymbol{X}_i \equiv \left[\begin{array}{c}\boldsymbol{X}_{i1}\\\vdots\\\boldsymbol{X}_{in_i}\end{array}\right] \quad \text{and put} \quad \boldsymbol{y} \equiv \left[\begin{array}{c}\boldsymbol{y}_1\\\vdots\\\boldsymbol{y}_m\end{array}\right] \quad \text{and} \quad \boldsymbol{X} \equiv \left[\begin{array}{c}\boldsymbol{X}_1\\\vdots\\\boldsymbol{X}_m\end{array}\right].$$

Then set

$$\boldsymbol{Z} \equiv \underset{1 \le i \le m}{\text{blockdiag}}\left[\underset{1 \le j \le n_i}{\text{stack}}(\boldsymbol{Z}_{ij}^{\text{L1}})\ \underset{1 \le j \le n_i}{\text{blockdiag}}(\boldsymbol{Z}_{ij}^{\text{L2}})\right], \quad \boldsymbol{u} \equiv \underset{1 \le i \le m}{\text{stack}}\left(\left[\begin{array}{c}\boldsymbol{u}_i^{\text{L1}}\\\boldsymbol{u}_{i1}^{\text{L2}}\\\vdots\\\boldsymbol{u}_{in_i}^{\text{L2}}\end{array}\right]\right),$$

$$\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}], \quad \boldsymbol{D}_{\text{BLUP}} \equiv \left[\begin{array}{cc}\boldsymbol{O} & \boldsymbol{O}\\\boldsymbol{O} & \underset{1 \le i \le m}{\text{blockdiag}}\left[\begin{array}{cc}(\boldsymbol{\Sigma}^{\text{L1}})^{-1} & \boldsymbol{O}\\\boldsymbol{O} & \boldsymbol{I}_{n_i} \otimes (\boldsymbol{\Sigma}^{\text{L2}})^{-1}\end{array}\right]\end{array}\right] \text{ and } \boldsymbol{R}_{\text{BLUP}} \equiv \sigma^2\boldsymbol{I}. \quad (35)$$

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

$$\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}\\\widehat{\boldsymbol{u}}\end{array}\right] = (\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1}\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{y}$$

$$\text{and} \quad \text{Cov}\left(\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}\\\widehat{\boldsymbol{u}} - \boldsymbol{u}\end{array}\right]\right) = (\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1}.$$

As in the two-level case, our goal is streamlined computation of $[\widehat{\boldsymbol{\beta}}^T\ \widehat{\boldsymbol{u}}^T]^T$ and the following covariance matrix sub-blocks:

$$\text{Cov}(\widehat{\boldsymbol{\beta}}), \quad \text{Cov}(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}}), \quad \text{Cov}(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}}),$$

$$E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}})^T\}, \quad E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}})^T\} \quad \text{and} \quad E\{(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}})(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}})^T\} \quad (36)$$

for $1 \le i \le m$ and $1 \le j \le n_i$. Analogously to (20), each of these matrices are sub-blocks of $(\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}})^{-1}$ with positions corresponding to the non-zero sub-blocks of $\boldsymbol{C}^T\boldsymbol{R}_{\text{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{BLUP}}$.

Result 5 shows that the required solutions can be embedded within the three-level sparse matrix least squares infrastructure.

Result 5. *Computation of $[\widehat{\boldsymbol{\beta}}^T \quad \widehat{\boldsymbol{u}}^T]^T$ and each of the sub-blocks of $\mathrm{Cov}([\widehat{\boldsymbol{\beta}}^T \quad (\widehat{\boldsymbol{u}} - \boldsymbol{u})^T]^T)$ listed in (36) are expressible as the three-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 illustrated by (13) are, for $1 \le i \le m$, $1 \le j \le n_i$:

$$\boldsymbol{b}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{y}_{ij} \\ 0 \\ 0 \end{bmatrix}, \; \boldsymbol{B}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{X}_{ij} \\ \boldsymbol{O} \\ \boldsymbol{O} \end{bmatrix}, \; \dot{\boldsymbol{B}}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_{ij}^{\mathrm{L1}} \\ n_i^{-1/2}(\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1/2} \\ \boldsymbol{O} \end{bmatrix} \; and \; \ddot{\boldsymbol{B}}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_{ij}^{\mathrm{L2}} \\ \boldsymbol{O} \\ (\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1/2} \end{bmatrix},$$

with each of these matrices having $\tilde{o}_{ij} = o_{ij} + q_1 + q_2$ rows. The solutions are

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{x}_1, \quad \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{A}^{11},$$

$$\widehat{\boldsymbol{u}}_i^{\mathrm{L1}} = \boldsymbol{x}_{2,i}, \; E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i^{\mathrm{L1}} - \boldsymbol{u}_i^{\mathrm{L1}})^T\} = \boldsymbol{A}^{12,i}, \; \mathrm{Cov}(\widehat{\boldsymbol{u}}_i^{\mathrm{L1}} - \boldsymbol{u}_i^{\mathrm{L1}}) = \boldsymbol{A}^{22,i}, \; for \; 1 \le i \le m.$$

$$\widehat{\boldsymbol{u}}_i^{\mathrm{L2}} = \boldsymbol{x}_{2,ij}, \; E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_{ij}^{\mathrm{L2}} - \boldsymbol{u}_{ij}^{\mathrm{L2}})^T\} = \boldsymbol{A}^{12,ij}, \; E\{\widehat{\boldsymbol{u}}_i^{\mathrm{L1}}(\widehat{\boldsymbol{u}}_{ij}^{\mathrm{L2}} - \boldsymbol{u}_{ij}^{\mathrm{L2}})^T\} = \boldsymbol{A}^{12,i,j}, \; \mathrm{Cov}(\widehat{\boldsymbol{u}}_{ij}^{\mathrm{L2}} - \boldsymbol{u}_{ij}^{\mathrm{L2}}) = \boldsymbol{A}^{22,ij}$$

for $1 \le i \le m$, $1 \le j \le n_i$.

Algorithm 10 is an immediate consequence of Result 5.

Algorithm 10 *QR-decomposition-based streamlined algorithm for obtaining best linear unbiased predictions and corresponding covariance matrix components for the three-level linear mixed model.*

Inputs: $\boldsymbol{y}_{ij}(o_{ij} \times 1), \boldsymbol{X}_{ij}(o_{ij} \times p), \boldsymbol{Z}_{ij}^{\text{L1}}(o_{ij} \times q_1), \boldsymbol{Z}_{ij}^{\text{L2}}(o_{ij} \times q_2),$
$\quad\quad 1 \le i \le m, \ 1 \le j \le n_i; \quad \sigma^2 > 0, \quad \boldsymbol{\Sigma}^{\text{L1}}(q_1 \times q_1), \ \boldsymbol{\Sigma}^{\text{L2}}(q_2 \times q_2),$
$\quad\quad$ both symmetric and positive definite.

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

$\quad\quad$ For $j = 1, \ldots, n_i$:

$$\boldsymbol{b}_{ij} \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{y}_{ij} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \ ; \ \ \boldsymbol{B}_{ij} \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{X}_{ij} \\ \boldsymbol{O} \\ \boldsymbol{O} \end{bmatrix}$$

$$\dot{\boldsymbol{B}}_{ij} \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_{ij}^{\text{L1}} \\ n_i^{-1/2}(\boldsymbol{\Sigma}^{\text{L1}})^{-1/2} \\ \boldsymbol{O} \end{bmatrix} \ ; \ \ \ddot{\boldsymbol{B}}_{ij} \longleftarrow \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_{ij}^{\text{L2}} \\ \boldsymbol{O} \\ (\boldsymbol{\Sigma}^{\text{L2}})^{-1/2} \end{bmatrix},$$

$\mathcal{S}_7 \longleftarrow$ SolveThreeLevelSparseLeastSquares
$\quad\quad \left(\left\{ (\boldsymbol{b}_{ij}, \boldsymbol{B}_{ij}, \dot{\boldsymbol{B}}_{ij}, \ddot{\boldsymbol{B}}_{ij}) : 1 \le i \le m, 1 \le j \le n_i \right\} \right)$

$\widehat{\boldsymbol{\beta}} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_7$ $\ ; \ $ $\text{Cov}(\widehat{\boldsymbol{\beta}}) \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_7$

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

$\quad\quad \widehat{\boldsymbol{u}}_i^{\text{L1}} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_7$ $\ ; \ $ $\text{Cov}(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}}) \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_7$

$\quad\quad E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_1$

$\quad\quad$ For $j = 1, \ldots, n_i$:

$\quad\quad\quad\quad \widehat{\boldsymbol{u}}_{ij}^{\text{L2}} \longleftarrow \boldsymbol{x}_{2,ij}$ component of $\mathcal{S}_7$ $\ ; \ $ $\text{Cov}(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}}) \longleftarrow \boldsymbol{A}^{22,ij}$ component of $\mathcal{S}_7$

$\quad\quad\quad\quad E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}})^T\} \longleftarrow \boldsymbol{A}^{12,ij}$ component of $\mathcal{S}_7$

$\quad\quad\quad\quad E\{(\widehat{\boldsymbol{u}}_{ij}^{\text{L1}} - \boldsymbol{u}_{ij}^{\text{L1}})(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}})^T\} \longleftarrow \boldsymbol{A}^{12,i,j}$ component of $\mathcal{S}_7$

Outputs: $\widehat{\boldsymbol{\beta}}, \text{Cov}(\widehat{\boldsymbol{\beta}}), \left\{ \left(\widehat{\boldsymbol{u}}_i^{\text{L1}}, \text{Cov}(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}}), E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_i^{\text{L1}} - \boldsymbol{u}_i^{\text{L1}})^T\} \right) : 1 \le i \le m \right\},$

$\quad\quad \left\{ \left(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}}, \text{Cov}(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}}), E\{\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}})^T\}, E\{(\widehat{\boldsymbol{u}}_{ij}^{\text{L1}} - \boldsymbol{u}_{ij}^{\text{L1}})(\widehat{\boldsymbol{u}}_{ij}^{\text{L2}} - \boldsymbol{u}_{ij}^{\text{L2}})^T\} \right) :$
$\quad\quad 1 \le i \le m, \ 1 \le j \le n_i \right\}$

5.2 Mean Field Variational Bayes

A Bayesian version of the three-level linear mixed model treated in the previous subsection is

$$\boldsymbol{y}_{ij}|\boldsymbol{\beta}, \boldsymbol{u}_i^{\text{L1}}, \boldsymbol{u}_{ij}^{\text{L2}}, \sigma^2 \overset{\text{ind.}}{\sim} N(\boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{u}_i^{\text{L1}} + \boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{u}_{ij}^{\text{L2}}, \sigma^2\,\boldsymbol{I}),$$

$$\left[\begin{array}{c} \boldsymbol{u}_i^{\text{L1}} \\ \boldsymbol{u}_{ij}^{\text{L2}} \end{array}\right] \Big| \boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}} \overset{\text{ind.}}{\sim} N\left(\left[\begin{array}{c} \boldsymbol{0} \\ \boldsymbol{0} \end{array}\right], \left[\begin{array}{cc} \boldsymbol{\Sigma}^{\text{L1}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}^{\text{L2}} \end{array}\right]\right), \quad 1 \leq i \leq m,\ 1 \leq j \leq n_i,$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \quad \sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(\nu_{\sigma^2}, 1/a_{\sigma^2}),$$

$$a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\sigma^2}s_{\sigma^2}^2)), \tag{37}$$

$$\boldsymbol{\Sigma}^{\text{L1}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}^{\text{L1}}} + 2q_1 - 2, (\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}})^{-1}\big),$$

$$\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \{\nu_{\boldsymbol{\Sigma}^{\text{L1}}}\text{diag}(s_{\boldsymbol{\Sigma}^{\text{L1}},1}^2, \ldots, s_{\boldsymbol{\Sigma}^{\text{L1}},q_1}^2)\}^{-1}),$$

$$\boldsymbol{\Sigma}^{\text{L2}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2 - 2, (\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})^{-1}\big),$$

$$\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \{\nu_{\boldsymbol{\Sigma}^{\text{L2}}}\text{diag}(s_{\boldsymbol{\Sigma}^{\text{L2}},1}^2, \ldots, s_{\boldsymbol{\Sigma}^{\text{L2}},q_2}^2)\}^{-1}).$$

where hyperparameters such as $\nu_{\boldsymbol{\Sigma}^{\text{L1}}} > 0$ and $s_{\boldsymbol{\Sigma}^{\text{L1}},1}, \ldots, s_{\boldsymbol{\Sigma}^{\text{L1}},q_1} > 0$ are defined analogously to the two-level case.

The minimal mean field restriction needed for a tractable variational inference algorithm is

$$\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, a_{\sigma^2}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}, \sigma^2, \boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}}|\boldsymbol{y}) \approx \mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, a_{\sigma^2}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})\,\mathfrak{q}(\sigma^2, \boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}}). \tag{38}$$

The optimal q-densities have forms analogous to those given in (23) but with

$$\mathfrak{q}^*(\boldsymbol{\Sigma}^{\text{L1}}) \text{ an Inverse-G-Wishart}\Big(G_{\text{full}}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L1}})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L1}})}\Big)$$

density function. A similar result holds for $\mathfrak{q}^*(\boldsymbol{\Sigma}^{\text{L2}})$.

Result 6 is the three-level analog of Result 2 in that it provides a link between the three-level sparse matrix least squares problems and updates for $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ and the important sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$.

Result 6. *The mean field variational Bayes updates of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ and each of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ in (26) are expressible as a three-level sparse matrix least squares problem of the form:*

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

where $\boldsymbol{b}$ and the non-zero sub-blocks $\boldsymbol{B}$, according to the notation given by (15), are for $1 \leq j \leq n_i$, $1 \leq i \leq m$:

$$\boldsymbol{b}_{ij} \equiv \left[\begin{array}{c} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{y}_{ij} \\ \left(\sum_{i=1}^m n_i\right)^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{array}\right], \quad \boldsymbol{B}_{ij} \equiv \left[\begin{array}{c} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{X}_{ij} \\ \left(\sum_{i=1}^m n_i\right)^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \\ \boldsymbol{O} \end{array}\right],$$

$$
\dot{\boldsymbol{B}}_{ij} \equiv
\begin{bmatrix}
\mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{Z}_{ij}^{\text{L1}} \\
\boldsymbol{O} \\
n_i^{-1/2}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})}\right)^{1/2} \\
\boldsymbol{O}
\end{bmatrix}
\quad and \quad
\ddot{\boldsymbol{B}}_{ij} \equiv
\begin{bmatrix}
\mu_{\mathfrak{q}(1/\sigma^2)}^{1/2}\boldsymbol{Z}_{ij}^{\text{L2}} \\
\boldsymbol{O} \\
\boldsymbol{O} \\
\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}\right)^{1/2}
\end{bmatrix}
$$

with each of these matrices having $\tilde{o}_{ij} = o_{ij} + p + q_1 + q_2$ rows. The solutions are

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

$$
\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} = \boldsymbol{x}_{2,i}, \ \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} = \boldsymbol{A}^{22,i}, \ E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\} = \boldsymbol{A}^{12,i} \ for \ 1 \le i \le m
$$

and

$$
\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})} = \boldsymbol{x}_{2,ij}, \ \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})} = \boldsymbol{A}^{22,ij}, \ E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\} = \boldsymbol{A}^{12,ij},
$$

$$
E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\} = \boldsymbol{A}^{12,i,j} \ for \ 1 \le i \le m, \ 1 \le j \le n_i.
$$

Algorithm 11 provides a streamlined mean field variational Bayes algorithm for approximate fitting and inference for (37). An explicit streamlined expression for the stopping criterion, $\log\{\mathfrak{p}(\boldsymbol{y}; \mathfrak{q})\}$, is given in Section A.15 of the appendix. We are not aware of any previously published variational inference algorithms that achieve streamlined inference for mixed models with three-level random effects.

5.3 Variational Message Passing

For studying the variational message passing alternative we first note that the joint density function of all of the random variables and random vectors in the Bayesian three-level Gaussian response linear mixed model (37) can be factorized as follows:

$$
\mathfrak{p}(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, \boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}}, a_{\sigma^2}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}) = \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)\mathfrak{p}(\sigma^2|a_{\sigma^2})\mathfrak{p}(a_{\sigma^2})
$$
$$
\times \mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}})\mathfrak{p}(\boldsymbol{\Sigma}^{\text{L1}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}})\mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}})\mathfrak{p}(\boldsymbol{\Sigma}^{\text{L2}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})\mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}).
$$

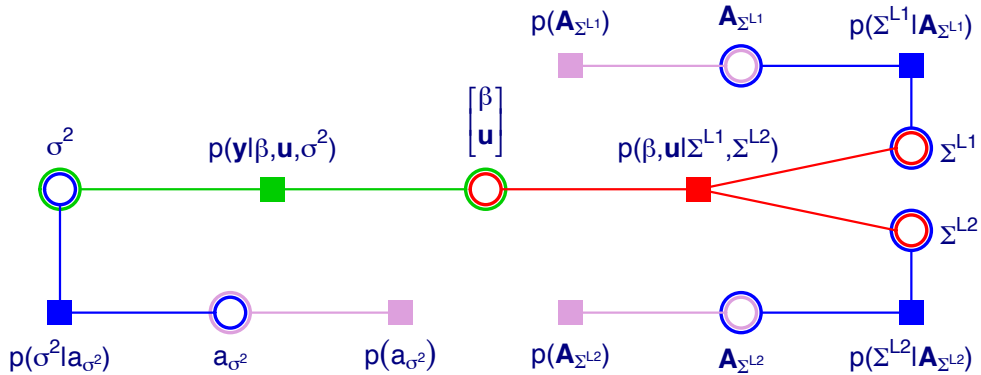Figure 6 provides the relevant factor graph with color-coding of fragment types.

Figure 6: *Factor graph representation of the Bayesian three-level Gaussian response linear mixed model (37).*

As with the two-level case, each of these fragments in Figure 6 appear in Section 4.1 of Wand (2017). To achieve streamlined variational message passing for three-level random

Algorithm 11 *QR-decomposition-based streamlined algorithm for obtaining mean field variational Bayes approximate posterior density functions for the parameters in the three-level linear mixed model (37) with product density restriction (38). The algorithm description requires more than one page and is continued on a subsequent page.*

Data Inputs: $\boldsymbol{y}_{ij}(o_{ij} \times 1)$, $\boldsymbol{X}_{ij}(o_{ij} \times p)$, $\boldsymbol{Z}_{ij}^{\mathrm{L1}}(o_{ij} \times q_1), \boldsymbol{Z}_{ij}^{\mathrm{L2}}(o_{ij} \times q_2)$, $1 \le i \le m$, $1 \le j \le n_i$.

Hyperparameter Inputs: $\boldsymbol{\mu_\beta}(p \times 1)$, $\boldsymbol{\Sigma_\beta}(p \times p)$ symmetric and positive definite,
$$s_{\sigma^2}, \nu_{\sigma^2}, s_{\boldsymbol{\Sigma}^{\mathrm{L1}},1}, \ldots, s_{\boldsymbol{\Sigma}^{\mathrm{L1}},q_1}, \nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}}, s_{\boldsymbol{\Sigma}^{\mathrm{L2}},1}, \ldots, s_{\boldsymbol{\Sigma}^{\mathrm{L2}},q_2}, \nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}} > 0$$

Initialize: $\mu_{\mathsf{q}(1/\sigma^2)} > 0$, $\mu_{\mathsf{q}(1/a_{\sigma^2})} > 0$, $\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})}(q_1 \times q_1)$, $\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})}(q_2 \times q_2)$,

$\boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}^{-1})}(q_1 \times q_1)$, $\boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}^{-1})}(q_2 \times q_2)$ symmetric and positive definite,

$$\xi_{\mathsf{q}(\sigma^2)} \longleftarrow \nu_{\sigma^2} + \sum_{i=1}^{m}\sum_{j=1}^{n_i} o_{ij} \ ; \ \xi_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} \longleftarrow \nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}} + 2q_1 - 2 + m \ ; \ \xi_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} \longleftarrow \nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}} + 2q_2 - 2 + \sum_{i=1}^{m} n_i$$

$$\xi_{\mathsf{q}(a_{\sigma^2})} \longleftarrow \nu_{\sigma^2} + 1 \ ; \ \xi_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})} \longleftarrow \nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}} + q_1 \ ; \ \xi_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})} \longleftarrow \nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}} + q_2$$

Cycle:

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

For $j = 1, \ldots, n_i$:

$$\boldsymbol{b}_{ij} \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma_\varepsilon^2)}^{1/2}\boldsymbol{y}_{ij} \\ \left(\sum_{i=1}^{m} n_i\right)^{-1/2}\boldsymbol{\Sigma_\beta}^{-1/2}\boldsymbol{\mu_\beta} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \ ; \ \boldsymbol{B}_{ij} \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma_\varepsilon^2)}^{1/2}\boldsymbol{X}_{ij} \\ \left(\sum_{i=1}^{m} n_i\right)^{-1/2}\boldsymbol{\Sigma_\beta}^{-1/2} \\ \boldsymbol{O} \\ \boldsymbol{O} \end{bmatrix},$$

$$\dot{\boldsymbol{B}}_{ij} \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma_\varepsilon^2)}^{1/2}\boldsymbol{Z}_{ij}^{\mathrm{L1}} \\ \boldsymbol{O} \\ n_i^{-1/2}\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})}^{1/2} \\ \boldsymbol{O} \end{bmatrix} \ ; \ \ddot{\boldsymbol{B}}_{ij} \longleftarrow \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma_\varepsilon^2)}^{1/2}\boldsymbol{Z}_{ij}^{\mathrm{L2}} \\ \boldsymbol{O} \\ \boldsymbol{O} \\ \boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})}^{1/2} \end{bmatrix}$$

$\mathcal{S}_8 \longleftarrow \textsc{SolveThreeLevelSparseLeastSquares}$
$$\left(\left\{(\boldsymbol{b}_{ij}, \boldsymbol{B}_{ij}, \dot{\boldsymbol{B}}_{ij}, \ddot{\boldsymbol{B}}_{ij}) : 1 \le i \le m, 1 \le j \le n_i\right\}\right)$$

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_8$

$\lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \mu_{\mathsf{q}(1/a_{\sigma^2})}$; $\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} \longleftarrow \boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}^{-1})}$; $\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} \longleftarrow \boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}^{-1})}$

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

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_8$

$E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_8$

$\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} \longleftarrow \boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} + \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}^T + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}$

For $j = 1, \ldots, n_i$:

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})} \longleftarrow \boldsymbol{x}_{2,ij}$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})} \longleftarrow \boldsymbol{A}^{22,ij}$ component of $\mathcal{S}_8$

$E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\} \longleftarrow \boldsymbol{A}^{12,ij}$ component of $\mathcal{S}_8$

$E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\} \longleftarrow \boldsymbol{A}^{12,i,j}$ component of $\mathcal{S}_8$

$\lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \lambda_{\mathsf{q}(\sigma^2)} + \left\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} - \boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\right\|^2$

$\lambda_{\mathsf{q}(\sigma^2)} \longleftarrow \lambda_{\mathsf{q}(\sigma^2)} + \mathrm{tr}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}) + \mathrm{tr}\{(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\}$

continued on a subsequent page . . .

Algorithm 11 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\lambda_{\mathfrak{q}(\sigma^2)} \longleftarrow \lambda_{\mathfrak{q}(\sigma^2)} + \mathrm{tr}((\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T \boldsymbol{Z}_{ij}^{\mathrm{L2}} \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})$$

$$\lambda_{\mathfrak{q}(\sigma^2)} \longleftarrow \lambda_{\mathfrak{q}(\sigma^2)} + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T \boldsymbol{X}_{ij} E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^T\}\big]$$

$$\lambda_{\mathfrak{q}(\sigma^2)} \longleftarrow \lambda_{\mathfrak{q}(\sigma^2)} + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T \boldsymbol{X}_{ij} E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}\big]$$

$$\lambda_{\mathfrak{q}(\sigma^2)} \longleftarrow \lambda_{\mathfrak{q}(\sigma^2)} + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T \boldsymbol{Z}_{ij}^{\mathrm{L2}} E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}\big]$$

$$\mu_{\mathfrak{q}(1/\sigma^2)} \longleftarrow \xi_{\mathfrak{q}(\sigma^2)}/\lambda_{\mathfrak{q}(\sigma^2)}$$

$$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} \longleftarrow (\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} - q_1 + 1)\,\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})}^{-1} \quad;\quad \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} \longleftarrow (\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} - q_2 + 1)\,\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}^{-1}$$

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

$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})} \longleftarrow \mathrm{diag}\{\mathrm{diagonal}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})}\big)\} + \{\nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}}\,\mathrm{diag}(s_{\boldsymbol{\Sigma}^{\mathrm{L1}},\,1}^2,\dots,s_{\boldsymbol{\Sigma}^{\mathrm{L1}},\,q_1}^2)\}^{-1}$$

$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})} \longleftarrow \mathrm{diag}\{\mathrm{diagonal}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})}\big)\} + \{\nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}}\,\mathrm{diag}(s_{\boldsymbol{\Sigma}^{\mathrm{L2}},\,1}^2,\dots,s_{\boldsymbol{\Sigma}^{\mathrm{L2}},\,q_2}^2)\}^{-1}$$

$$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}^{-1})} \longleftarrow \xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})}\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})}^{-1} \quad;\quad \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}^{-1})} \longleftarrow \xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})}\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})}^{-1}.$$

until the increase in $\log\{\underline{p}(\boldsymbol{y};\mathfrak{q})\}$ is negligible.

Outputs: $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}, \big\{\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}, E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^T\}\big) : 1 \le i \le m\big\},$

$\qquad \big\{\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}, E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\},$

$\qquad E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}\big) : 1 \le i \le m, 1 \le j \le n_i\big\},$

$\qquad \xi_{\mathfrak{q}(\sigma^2)}, \lambda_{\mathfrak{q}(\sigma^2)}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}$

---

effects models we require tailored versions of the Gaussian likelihood fragment updates and Gaussian penalization fragment updates. These are provided in the next two subsections as Algorithms 13 and 14. However, they each rely on the THREELEVELNATURALTOCOMMONPARAMETERS algorithm, which is listed as Algorithm 12.

5.4 Streamlined Gaussian Likelihood Fragment Updates

Streamlined updating for the Gaussian likelihood fragment with three-level random effects structure is analogous to the two-level case discussed in Section 4.4. The relevant factor is shown in green in Figure 6. The message from the likelihood factor to the vector of fixed and random effects instead has the form

Algorithm 12 *The* THREELEVELNATURALTOCOMMONPARAMETERS *algorithm. The algorithm description requires more than one page and is continued on a subsequent page.*

Inputs: $p, q_1, q_2, m, \{n_i : 1 \le i \le m\}, \boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$

$\boldsymbol{\omega}_{26} \longleftarrow$ first p entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$

$\boldsymbol{\omega}_{27} \longleftarrow$ next $\frac{1}{2}p(p+1)$ entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$; $\boldsymbol{\Omega}_{28} \longleftarrow -2\text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\omega}_{27})$

$i_{\text{stt}} \longleftarrow p + \frac{1}{2}p(p+1) + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_1 - 1$;

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

$\quad\quad \boldsymbol{\omega}_{29i} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + \frac{1}{2}q_1(q_1+1) - 1$

$\quad\quad \boldsymbol{\omega}_{30} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq_1 - 1$

$\quad\quad \boldsymbol{\omega}_{31} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_1 - 1$

$\quad\quad \boldsymbol{\Omega}_{32i} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\omega}_{30})$; $\boldsymbol{\Omega}_{33i} \longleftarrow -\text{vec}_{p \times q_1}^{-1}(\boldsymbol{\omega}_{31})$

$i_{\text{end}} \longleftarrow i_{\text{end}} - q_1 + q_2$

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

$\quad\quad$ For $j = 1, \dots, n_i$:

$\quad\quad\quad\quad \boldsymbol{\omega}_{34ij} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + \frac{1}{2}q_2(q_2+1) - 1$

$\quad\quad\quad\quad \boldsymbol{\omega}_{35} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq_2 - 1$

$\quad\quad\quad\quad \boldsymbol{\omega}_{36} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_1 q_2 - 1$

$\quad\quad\quad\quad \boldsymbol{\omega}_{37} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive

$\quad\quad\quad\quad i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_2 - 1$

$\quad\quad\quad\quad \boldsymbol{\Omega}_{38ij} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\omega}_{35})$; $\boldsymbol{\Omega}_{39ij} \longleftarrow -\text{vec}_{p \times q_2}^{-1}(\boldsymbol{\omega}_{36})$

$\quad\quad\quad\quad \boldsymbol{\Omega}_{40ij} \longleftarrow -\text{vec}_{q_1 \times q_2}^{-1}(\boldsymbol{\omega}_{37})$

$\mathcal{S}_8 \longleftarrow$ SOLVETHREELEVELSPARSEMATRIX$\Big(\boldsymbol{\omega}_{26}, \boldsymbol{\Omega}_{28}, \big\{(\boldsymbol{\omega}_{29i}, \boldsymbol{\Omega}_{32i}, \boldsymbol{\Omega}_{33i}) : 1 \le i \le m,$

$$(\boldsymbol{\omega}_{34ij}, \boldsymbol{\Omega}_{38ij}, \boldsymbol{\Omega}_{39ij}, \boldsymbol{\Omega}_{40ij}) : 1 \le i \le m, 1 \le j \le n_i \big\}\Big)$$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_8$

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

$\quad\quad \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_8$,

$\quad\quad E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})\}(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_8$

$\quad\quad$ *continued on a subsequent page* ...

Algorithm 12 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

For $j = 1, \ldots, n_i$:

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})} \longleftarrow \boldsymbol{x}_{2,ij}$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})} \longleftarrow \boldsymbol{A}^{22,ij}$ component of $\mathcal{S}_8$

$E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})\}(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})})^T\} \longleftarrow \boldsymbol{A}^{12,ij}$ component of $\mathcal{S}_8$

$E_{\mathsf{q}}\{(\boldsymbol{u}_i^{L1} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{L1})})\}(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})})^T\} \longleftarrow \boldsymbol{A}^{12,i,j}$ component of $\mathcal{S}_8$

Outputs: $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}, \left\{\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{L1})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{L1})}, E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}\}(\boldsymbol{u}_i^{L1} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{L1})})^T\}\right) : 1 \le i \le m\right\}$,

$\left\{\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})}, \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})}, E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}\}(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})})^T\},\right.\right.$

$\left.\left. E_{\mathsf{q}}\{(\boldsymbol{u}_i^{L1} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{L1})})\}(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})})^T\}\right) : 1 \le i \le m, \ 1 \le j \le n_i\right\}$

$$m_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) =$$

$$\exp\left\{\left[\begin{array}{c} \left[\begin{array}{c} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1 \le i \le m}{\text{stack}}\left[\begin{array}{c} \boldsymbol{u}_i^{L1} \\ \text{vech}(\boldsymbol{u}_i^{L1}(\boldsymbol{u}_i^{L1})^T) \\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_i^{L1})^T) \end{array}\right] \\ \underset{1 \le i \le m}{\text{stack}}\left[\underset{1 \le j \le n_i}{\text{stack}}\left[\begin{array}{c} \boldsymbol{u}_{ij}^{L2} \\ \text{vech}(\boldsymbol{u}_{ij}^{L2}(\boldsymbol{u}_{ij}^{L2})^T) \\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{L1})^T) \\ \text{vec}(\boldsymbol{u}_{ij}^{L1}(\boldsymbol{u}_{ij}^{L2})^T) \end{array}\right]\right] \end{array}\right]^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}\right\} \quad (39)$$

and we assume that $m_{(\boldsymbol{\beta},\boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u})$ is in the same exponential family. Result 7 points the way to streamlining the fragment updates in the three-level case. Its derivation is given in Section A.17.

Result 7. *The variational message passing updates of the quantities $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}$, $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{L1})}$, $1 \le i \le m$, $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{L2})}$, $1 \le i \le m$, $1 \le j \le n_i$, and the sub-blocks of $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ listed in (36) with q-density expectations with respect to the normalization of*

$$m_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\, m_{(\boldsymbol{\beta},\boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u})$$

are expressible as a three-level sparse matrix problem with

$$\boldsymbol{A} =$$
$$-2\left[\begin{array}{cc} \text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\eta}_{1,2}) & \left(\frac{1}{2}\underset{1 \le i \le m}{\text{stack}}\left[\text{vec}_{p \times q_1}^{-1}(\boldsymbol{\eta}_{2,3,i})^T \underset{1 \le j \le n_i}{\text{stack}}\{\text{vec}_{p \times q_2}^{-1}(\boldsymbol{\eta}_{3,3,ij})^T\}\right]\right)^T \\ \begin{array}{c}\frac{1}{2}\underset{1 \le i \le m}{\text{stack}}\left[\text{vec}_{p \times q_1}^{-1}(\boldsymbol{\eta}_{2,3,i})^T \\ \underset{1 \le j \le n_i}{\text{stack}}\{\text{vec}_{p \times q_2}^{-1}(\boldsymbol{\eta}_{3,3,ij})^T\}\right]\end{array} & \underset{1 \le i \le m}{\text{blockdiag}}\left[\begin{array}{cc} \text{vec}^{-1}(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\eta}_{2,2,i}) & \left[\frac{1}{2}\underset{1 \le j \le n_i}{\text{stack}}\{\text{vec}_{q_1 \times q_2}^{-1}(\boldsymbol{\eta}_{3,4,ij})^T\}\right]^T \\ \frac{1}{2}\underset{1 \le j \le n_i}{\text{stack}}\{\text{vec}_{q_1 \times q_2}^{-1}(\boldsymbol{\eta}_{3,4,ij})^T\} & \underset{1 \le j \le n_i}{\text{blockdiag}}\{\text{vec}^{-1}(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\eta}_{3,2,ij})\} \end{array}\right] \end{array}\right]$$

and

$$
\boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{\eta}_{1,1} \\[2mm] \underset{1\le i\le m}{\text{stack}}\,(\boldsymbol{\eta}_{2,1,i}) \\[2mm] \underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le j\le n_i}{\text{stack}}\,(\boldsymbol{\eta}_{3,1,ij})\right\} \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} \boldsymbol{\eta}_{1,1}\ (p\times 1) \\[1mm] \boldsymbol{\eta}_{1,2}\ (\frac{1}{2}p(p+1)\times 1) \\[2mm] \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \boldsymbol{\eta}_{2,1,i} & (q_1\times 1) \\ \boldsymbol{\eta}_{2,2,i} & (\frac{1}{2}q_1(q_1+1)\times 1) \\ \boldsymbol{\eta}_{2,3,i} & (pq_1\times 1) \end{bmatrix} \\[6mm] \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix} \boldsymbol{\eta}_{3,1,ij} & (q_2\times 1) \\ \boldsymbol{\eta}_{3,2,ij} & (\frac{1}{2}q_2(q_2+1)\times 1) \\ \boldsymbol{\eta}_{3,3,ij} & (pq_2\times 1) \\ \boldsymbol{\eta}_{3,4,ij} & (q_1q_2\times 1) \end{bmatrix}\end{bmatrix} \end{bmatrix}
$$

is the partitioning of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ that defines $\boldsymbol{\eta}_{1,1}$, $\boldsymbol{\eta}_{1,2}$, $\{(\boldsymbol{\eta}_{2,1,i},\boldsymbol{\eta}_{2,2,i},\boldsymbol{\eta}_{2,3,i}):1\le i\le m\}$ and $\{(\boldsymbol{\eta}_{3,1,ij},\boldsymbol{\eta}_{3,2,ij},\boldsymbol{\eta}_{3,3,ij},\boldsymbol{\eta}_{3,4,ij}):1\le i\le m, 1\le j\le n_i\}$. The solutions are $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}=\boldsymbol{x}_1$, $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}=\boldsymbol{A}^{11}$ and

$$
\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}=\boldsymbol{x}_{2,i},\ \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}=\boldsymbol{A}^{22,i},\ E_{\mathsf{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}=\boldsymbol{A}^{12,i}\ \text{for } 1\le i\le m
$$

and

$$
\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}=\boldsymbol{x}_{2,ij},\ \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}=\boldsymbol{A}^{22,ij},\ E_{\mathsf{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}=\boldsymbol{A}^{12,ij},
$$

$$
E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}=\boldsymbol{A}^{12,\,i,\,j}\ \text{for } 1\le i\le m,\ 1\le j\le n_i.
$$

The message from the likelihood factor to σ^2 has the form as in the two-level case, as given by (33). Streamlined Gaussian likelihood fragment updates for the messages from $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)$ to $(\boldsymbol{\beta},\boldsymbol{u})$ and σ^2 is encapsulated in Algorithm 13. Note its use of the notation defined by (34). Its justification is described in Section A.18.

5.5 Streamlined Gaussian Penalization Fragment Updates

Here we treat the Gaussian penalization fragment for three-level random effects structure. This fragment is shown in red in Figure 6. We assume that

$$
m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \quad \text{and} \quad m_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})
$$

are in the same exponential family. In other words, $m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})$ has the form given by the right-hand side of (39) but with natural parameter vector

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})} \quad \text{replaced by} \quad \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}.
$$

The fragment's other factor to stochastic node messages are

$$
m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,\boldsymbol{\Sigma}^{\text{L1}}}(\boldsymbol{\Sigma}^{\text{L1}})=\exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}^{\text{L1}}| \\ \text{vech}((\boldsymbol{\Sigma}^{\text{L1}})^{-1}) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,\boldsymbol{\Sigma}^{\text{L1}}}\right\}
$$

and

$$
m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,\boldsymbol{\Sigma}^{\text{L2}}}(\boldsymbol{\Sigma}^{\text{L2}})=\exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}^{\text{L2}}| \\ \text{vech}((\boldsymbol{\Sigma}^{\text{L2}})^{-1}) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\to\,\boldsymbol{\Sigma}^{\text{L2}}}\right\}.
$$

Streamlined updating of the three-level Gaussian penalization fragment is aided by Result 8:

Algorithm 13 *The inputs, updates and outputs of the matrix algebraic streamlined Gaussian likelihood fragment for three-level models. The algorithm description requires more than one page and is continued on a subsequent page.*

Data Inputs: $\boldsymbol{y}_{ij}(o_{ij} \times 1)$, $\boldsymbol{X}_{ij}(o_{ij} \times p)$, $\boldsymbol{Z}_{ij}^{\text{L1}}(o_{ij} \times q_1)$, $\boldsymbol{Z}_{ij}^{\text{L2}}(o_{ij} \times q_2)$, $1 \le i \le m$,
$\quad\quad 1 \le j \le n_i$.

Parameter Inputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to \sigma^2}$,
$\quad\quad \boldsymbol{\eta}_{\sigma^2 \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$

Updates:

$$\mu_{\mathsf{q}(1/\sigma^2)} \longleftarrow \left(\left(\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2} \right)_1 + 1 \right) \Big/ \left(\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2} \right)_2$$

$$\mathcal{S}_9 \longleftarrow \text{THREELEVELNATURALTOCOMMONPARAMETERS}\Big(p, q_1, q_2, m, \{n_i : 1 \le i \le m\},$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}\Big)$$

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}$ component of $\mathcal{S}_9$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}$ component of $\mathcal{S}_9$

$\boldsymbol{\omega}_{41} \longleftarrow \boldsymbol{0}_p$; $\boldsymbol{\omega}_{42} \longleftarrow \boldsymbol{0}_{\frac{1}{2}p(p+1)}$; $\boldsymbol{\omega}_{43} \longleftarrow 0$

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

$\quad\quad \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}$ component of $\mathcal{S}_9$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}$ component of $\mathcal{S}_9$

$\quad\quad E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})^T\} \longleftarrow E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}$

$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ component of $\mathcal{S}_9$

$\quad\quad \boldsymbol{\omega}_{44i} \longleftarrow \boldsymbol{0}_{q_1}$; $\boldsymbol{\omega}_{45i} \longleftarrow \boldsymbol{0}_{\frac{1}{2}q_1(q_1+1)}$; $\boldsymbol{\omega}_{46i} \longleftarrow \boldsymbol{0}_{p\,q_1}$

$\quad\quad$ For $j = 1, \ldots, n_i$:

$\quad\quad\quad\quad \boldsymbol{\omega}_{41} \longleftarrow \boldsymbol{\omega}_{41} + \boldsymbol{X}_{ij}^T \boldsymbol{y}_{ij}$; $\boldsymbol{\omega}_{42} \longleftarrow \boldsymbol{\omega}_{42} - \frac{1}{2}\boldsymbol{D}_p^T \text{vec}(\boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij})$

$\quad\quad\quad\quad \boldsymbol{\omega}_{44i} \longleftarrow \boldsymbol{\omega}_{44i} + (\boldsymbol{Z}_{ij}^{\text{L1}})^T \boldsymbol{y}_{ij}$; $\boldsymbol{\omega}_{45i} \longleftarrow \boldsymbol{\omega}_{45i} - \frac{1}{2}\boldsymbol{D}_{q_1}^T \text{vec}\big((\boldsymbol{Z}_{ij}^{\text{L1}})^T \boldsymbol{Z}_{ij}^{\text{L1}}\big)$

$\quad\quad\quad\quad \boldsymbol{\omega}_{46i} \longleftarrow \boldsymbol{\omega}_{46i} - \text{vec}\big(\boldsymbol{X}_{ij}^T \boldsymbol{Z}_{ij}^{\text{L1}}\big)$

$\quad\quad\quad\quad \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ component of $\mathcal{S}_9$

$\quad\quad\quad\quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ component of $\mathcal{S}_9$

$\quad\quad\quad\quad E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\} \longleftarrow E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}$

$\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ component of $\mathcal{S}_9$

$\quad\quad\quad\quad E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\} \longleftarrow$

$\quad\quad\quad\quad\quad\quad\quad E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}$ component of $\mathcal{S}_9$

$\quad\quad\quad\quad \omega_{43} \longleftarrow \omega_{43} - \frac{1}{2}\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})} - \boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\|^2$

$\quad\quad\quad\quad \omega_{43} \longleftarrow \omega_{43} - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}\boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij}) - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}(\boldsymbol{Z}_{ij}^{\text{L1}})^T \boldsymbol{Z}_{ij}^{\text{L1}})$

$\quad\quad\quad\quad\quad\quad - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\boldsymbol{Z}_{ij}^{L2T} \boldsymbol{Z}_{ij}^{L2})$

$\quad\quad\quad\quad\quad\quad - \text{tr}[\{(\boldsymbol{Z}_{ij}^{\text{L1}})^T \boldsymbol{X}_{ij} E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}]$

$\quad\quad\quad\quad\quad\quad - \text{tr}[\{(\boldsymbol{Z}_{ij}^{\text{L2}})^T \boldsymbol{X}_{ij} E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}]$

$\quad\quad\quad\quad\quad\quad - \text{tr}[\{(\boldsymbol{Z}_{ij}^{\text{L2}})^T \boldsymbol{Z}_{ij}^{\text{L1}} E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}]$

$\quad\quad$ *continued on a subsequent page* . . .

Algorithm 13 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}\longleftarrow\mu_{\mathsf{q}(1/\sigma^2)}\begin{bmatrix}\boldsymbol{\omega}_{41}\\\boldsymbol{\omega}_{42}\\\underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{\omega}_{44i}\\\boldsymbol{\omega}_{45i}\\\boldsymbol{\omega}_{46i}\end{bmatrix}\\\underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix}(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{y}_{ij}\\-\frac{1}{2}\boldsymbol{D}_{q_2}^T\mathrm{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\\-\mathrm{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\\-\mathrm{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,\sigma^2}\longleftarrow\begin{bmatrix}-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}o_{ij}\\\boldsymbol{\omega}_{43}\end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,\sigma^2}$.

Result 8. *The variational message passing updates of the quantities $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})},\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})},\ 1\le i\le m$, and $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})},\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})},\ 1\le i\le m,\ 1\le j\le n_i$, with q-density expectations with respect to the normalization of*

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\,m_{(\boldsymbol{\beta},\boldsymbol{u})\,\rightarrow\,\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})}(\boldsymbol{\beta},\boldsymbol{u})$$

are expressible as a three-level sparse matrix problem with

$\boldsymbol{A} =$

$$-2\begin{bmatrix}\mathrm{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\eta}_{1,2}) & \left(\frac{1}{2}\underset{1\le i\le m}{\text{stack}}\left[\mathrm{vec}_{p\times q_1}^{-1}(\boldsymbol{\eta}_{2,3,i})^T\underset{1\le j\le n_i}{\text{stack}}\{\mathrm{vec}_{p\times q_2}^{-1}(\boldsymbol{\eta}_{3,3,ij})^T\}\right]\right)^T\\[2em]\begin{bmatrix}\frac{1}{2}\underset{1\le i\le m}{\text{stack}}\left[\mathrm{vec}_{p\times q_1}^{-1}(\boldsymbol{\eta}_{2,3,i})^T\\\underset{1\le j\le n_i}{\text{stack}}\{\mathrm{vec}_{p\times q_2}^{-1}(\boldsymbol{\eta}_{3,3,ij})^T\}\right] & \underset{1\le i\le m}{\text{blockdiag}}\begin{bmatrix}\mathrm{vec}^{-1}(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\eta}_{2,2,i}) & \left[\frac{1}{2}\underset{1\le j\le n_i}{\text{stack}}\{\mathrm{vec}_{q_1\times q_2}^{-1}(\boldsymbol{\eta}_{3,4,ij})^T\}\right]^T\\[1em]\frac{1}{2}\underset{1\le j\le n_i}{\text{stack}}\{\mathrm{vec}_{q_1\times q_2}^{-1}(\boldsymbol{\eta}_{3,4,ij})^T\} & \underset{1\le j\le n_i}{\text{blockdiag}}\{\mathrm{vec}^{-1}(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\eta}_{3,2,ij})\}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

and

$$\boldsymbol{a}\equiv\begin{bmatrix}\boldsymbol{\eta}_{1,1}\\\underset{1\le i\le m}{\text{stack}}(\boldsymbol{\eta}_{2,1,i})\\\underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le j\le n_i}{\text{stack}}(\boldsymbol{\eta}_{3,1,ij})\right\}\end{bmatrix}\quad where\quad\begin{bmatrix}\boldsymbol{\eta}_{1,1}\ (p\times 1)\\\boldsymbol{\eta}_{1,2}\ (\frac{1}{2}p(p+1)\times 1)\\\underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{\eta}_{2,1,i}\ (q_1\times 1)\\\boldsymbol{\eta}_{2,2,i}\ (\frac{1}{2}q_1(q_1+1)\times 1)\\\boldsymbol{\eta}_{2,3,i}\ (pq_1\times 1)\end{bmatrix}\\\underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix}\boldsymbol{\eta}_{3,1,ij}\ (q_2\times 1)\\\boldsymbol{\eta}_{3,2,ij}\ (\frac{1}{2}q_2(q_2+1)\times 1)\\\boldsymbol{\eta}_{3,3,ij}\ (pq_2\times 1)\\\boldsymbol{\eta}_{3,4,ij}\ (q_1q_2\times 1)\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

is the partitioning of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ that defines $\boldsymbol{\eta}_{1,1},\boldsymbol{\eta}_{1,2},\{(\boldsymbol{\eta}_{2,1,i},\boldsymbol{\eta}_{2,2,i},\boldsymbol{\eta}_{2,3,i}):1\le i\le m\}$ and $\{(\boldsymbol{\eta}_{3,1,ij},\boldsymbol{\eta}_{3,2,ij},\boldsymbol{\eta}_{3,3,ij},\boldsymbol{\eta}_{3,4,ij}):1\le i\le m,1\le j\le n_i\}$. The solutions are

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}=\boldsymbol{x}_{2,i},\ \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}=\boldsymbol{A}^{22,i}\quad for\quad 1\le i\le m$$

and

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})} = \boldsymbol{x}_{2,ij}, \; \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})} = \boldsymbol{A}^{22,ij} \quad for \quad 1 \le i \le m, \; 1 \le j \le n_i.$$

Algorithm 14 provides the natural parameter vector updates for the three-level Gaussian penalization fragment based on Result 7. Note that natural parameter vectors containing a $\leftrightarrow$ in their subscript, such as $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}$, are defined by (34).

5.6 $\mathfrak{q}$-Density Determination After Variational Message Passing Convergence

The advice given in Section 4.6 for the two-level case extends straightforwardly to the three-level case. The main change is the need for the steps we need to first carry out:

$$\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$$

$$\mathcal{S}_{11} \longleftarrow \text{THREELEVELNATURALTOCOMMONPARAMETERS}\Big(p,q_1,q_2,m,\{n_i : 1 \le i \le m\},\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}\Big).$$

6   Timing Results

We obtained timing results for simulated data according to a version of model (21) for which both the fixed effects and random effects had dimension two, corresponding to random intercepts and slopes for a single continuous predictor which was generated from the Uniform distribution on the unit interval. The true parameter values were set to

$$\boldsymbol{\beta}_{\text{true}} = \left[\begin{array}{c} 0.58 \\ 1.98 \end{array} \right] \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\text{true}} = \left[\begin{array}{cc} 2.58 & 0.22 \\ 0.22 & 1.73 \end{array} \right]$$

and, throughout the study, the n_i values were generated uniformly on the set $\{30,\dots,60\}$. The study was run on a `MacBook Air` laptop with a 2.2 gigahertz processor and 8 gigabytes of random access memory. The number of mean field iterations was fixed at 50.

m	naïve	streamlined	naïve/streamlined
200	2.75 (0.0482)	0.035 (0.00000)	78.5
400	22.30 (0.2490)	0.070 (0.00148)	319.0
600	84.40 (0.4940)	0.108 (0.00445)	782.0
800	213.00 (0.9160)	0.143 (0.00445)	1490.0
1,000	427.00 (3.1000)	0.183 (0.00741)	2340.0

Table 1: *Median (median absolute deviation) of elapsed computing times in seconds for fitting model (21) naïvely versus with streamlining via Algorithm 6. The fourth column lists the ratios of the median computing times.*

The first phase of the study involved comparing the computational times of the streamlined Algorithm 6 with its naïve counterpart for which (24) was implemented directly. To allow for maximal speed, both approaches were implemented in the low-level language `Fortran 77`. The number of groups varied over $m \in \{200,400,600,800,1000\}$ and 100 replications were simulated for each value of m. For the most demanding $m = 1,000$ case the streamlined implementation had a median computing time of 0.183 seconds and a maximum of 0.354 seconds. By comparison, the naïve approach had a median computing time of 7 minutes and, for a few replications, took several hours. Because of such outliers

Algorithm 14 *The inputs, updates and outputs of the matrix algebraic streamlined Gaussian likelihood fragment for three-level models.*

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

Parameter Inputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to (\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to \boldsymbol{\Sigma}^{\text{L1}}}$,

$\boldsymbol{\eta}_{\boldsymbol{\Sigma}^{\text{L1}} \to \mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to \boldsymbol{\Sigma}^{\text{L2}}}$, $\boldsymbol{\eta}_{\boldsymbol{\Sigma}^{\text{L2}} \to \mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})}$

Updates:

$\omega_{47} \longleftarrow$ first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \leftrightarrow \boldsymbol{\Sigma}^{\text{L1}}}$

$\boldsymbol{\omega}_{48} \longleftarrow$ remaining entries of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \leftrightarrow \boldsymbol{\Sigma}^{\text{L1}}}$

$\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})} \longleftarrow \left\{\omega_{47} + \tfrac{1}{2}(q_1+1)\right\}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\omega}_{48}\right)\right\}^{-1}$

$\omega_{49} \longleftarrow$ first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \leftrightarrow \boldsymbol{\Sigma}^{\text{L2}}}$

$\boldsymbol{\omega}_{50} \longleftarrow$ remaining entries of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \leftrightarrow \boldsymbol{\Sigma}^{\text{L2}}}$

$\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})} \longleftarrow \left\{\omega_{49} + \tfrac{1}{2}(q_2+1)\right\}\left\{\text{vec}^{-1}\left(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\omega}_{50}\right)\right\}^{-1}$

$\mathcal{S}_{10} \longleftarrow$ THREELEVELNATURALTOCOMMONPARAMETERS
$$\left(p, q_1, q_2, m, \{n_i : 1 \le i \le m\}, \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}\right)$$

$\boldsymbol{\omega}_{51} \longleftarrow \boldsymbol{0}_{\frac{1}{2}q_1(q_1+1)}$; $\boldsymbol{\omega}_{52} \longleftarrow \boldsymbol{0}_{\frac{1}{2}q_2(q_2+1)}$

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

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}$ component of $\mathcal{S}_{10}$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}$ component of $\mathcal{S}_{10}$

$\boldsymbol{\omega}_{51} \longleftarrow \boldsymbol{\omega}_{51} - \tfrac{1}{2}\boldsymbol{D}_{q_1}^T\text{vec}\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}^T + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{L1}})}\right)$

For $j = 1, \ldots, n_i$:

$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})} \longleftarrow \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ component of $\mathcal{S}_{10}$; $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})} \longleftarrow \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ component of $\mathcal{S}_{10}$

$\boldsymbol{\omega}_{52} \longleftarrow \boldsymbol{\omega}_{52} - \tfrac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^T + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\right)$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to (\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\ \underset{1\le i\le m}{\text{stack}} \begin{bmatrix} \boldsymbol{0}_{q_1} \\ -\tfrac{1}{2}\boldsymbol{D}_{q_1}^T\text{vec}\left(\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})}\right) \\ \boldsymbol{0}_{pq_1} \end{bmatrix} \\ \underset{1\le i\le m}{\text{stack}} \left[\underset{1\le j\le n_i}{\text{stack}} \begin{bmatrix} \boldsymbol{0}_{q_2} \\ -\tfrac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}\left(\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}\right) \\ \boldsymbol{0}_{pq_2} \\ \boldsymbol{0}_{q_1 q_2} \end{bmatrix} \right] \end{bmatrix}$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to \boldsymbol{\Sigma}^{\text{L1}}} \longleftarrow \begin{bmatrix} -\tfrac{1}{2}m \\ \boldsymbol{\omega}_{51} \end{bmatrix} \quad ; \quad \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to \boldsymbol{\Sigma}^{\text{L2}}} \longleftarrow \begin{bmatrix} -\tfrac{1}{2}\sum_{i=1}^{m}n_i \\ \boldsymbol{\omega}_{52} \end{bmatrix}$$

Parameter Outputs: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to (\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to \boldsymbol{\Sigma}^{\text{L1}}}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) \to \boldsymbol{\Sigma}^{\text{L2}}}$

---

in the naïve computational times our summary of this first phase, given in Table 1, uses medians and median absolute deviations. As the number of groups increases into the sev-

eral hundreds we see that streamlined variational inference becomes thousands of times faster in terms of median performance.

The second phase of our timing study involved ramping up the number of groups into the tens of thousands and recording computational times for Algorithm 6. We used the geometric progression $m \in \{400, 1200, 3600, 10800, 32400\}$ and another 100 replications. Table 2 shows that the average computing times increase approximately linearly with m and only around 7 seconds are required for handling $m = 32,400$ groups.

$m = 400$	$m = 1,200$	$m = 3,600$	$m = 10,800$	$m = 32,400$
0.0781	0.2400	0.7140	2.30	6.980
(0.0122)	(0.0343)	(0.0806)	(0.270)	(0.857)

Table 2: *Average (standard deviation) of elapsed computing times in seconds for fitting model (21) with streamlining via Algorithm 6.*

In summary, the streamlined approach is vastly superior to naïve implementation in terms of speed and scales well to large data multilevel data situations.

7   Illustration for Data From a Large Longitudinal Perinatal Study

We now provide illustration for data from the Collaborative Perinatal Project, a large longitudinal perinatal health study that was run in the United States of America during 1959–1974 (e.g. Klebanoff, 2009). The data are publicly available from tue U.S. National Archives with identifier 606622. For our illustration in this section, which focuses on the first year of life, the number of infants followed longitudinally is 44,708 and the number of fields is 125,564. We do not perform a full-blown analysis of these data and eschew matters such as careful variable creation, model selection and intepretation. Instead we consider an illustrative Bayesian mixed model, with two-level random effects, and compare streamlined mean field variational Bayes and Markov chain Monte Carlo fits. Specifically, we consider the model

$$
\begin{aligned}
y_{ij}|\beta_0, \ldots, \beta_7, \sigma^2 &\overset{\text{ind.}}{\sim} N\Big(\beta_0 + u_{i0} + (\beta_1 + u_{i1})x_{1ij} + (\beta_1 + u_{i2})x_{2ij}^2 \\
&\qquad + \beta_3\, x_{3ij} + \ldots + \beta_7\, x_{7ij}, \sigma^2\Big),
\end{aligned}
$$

$$
\begin{bmatrix} u_{i0} \\ u_{i1} \\ u_{i2} \end{bmatrix} \Big| \boldsymbol{\Sigma} \overset{\text{ind.}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}), \text{ for } 1 \leq i \leq 44,708 \text{ and } 1 \leq j \leq n_i
$$

(40)

with priors

$$
\beta_0, \ldots, \beta_7 \overset{\text{ind.}}{\sim} N(0, 10^{10}), \quad \sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/a_{\sigma^2}),
$$

$$
a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 10^{-10}), \quad \boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, 6, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big),
$$

$$
\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, 2 \times 10^{-10}\boldsymbol{I}_3)
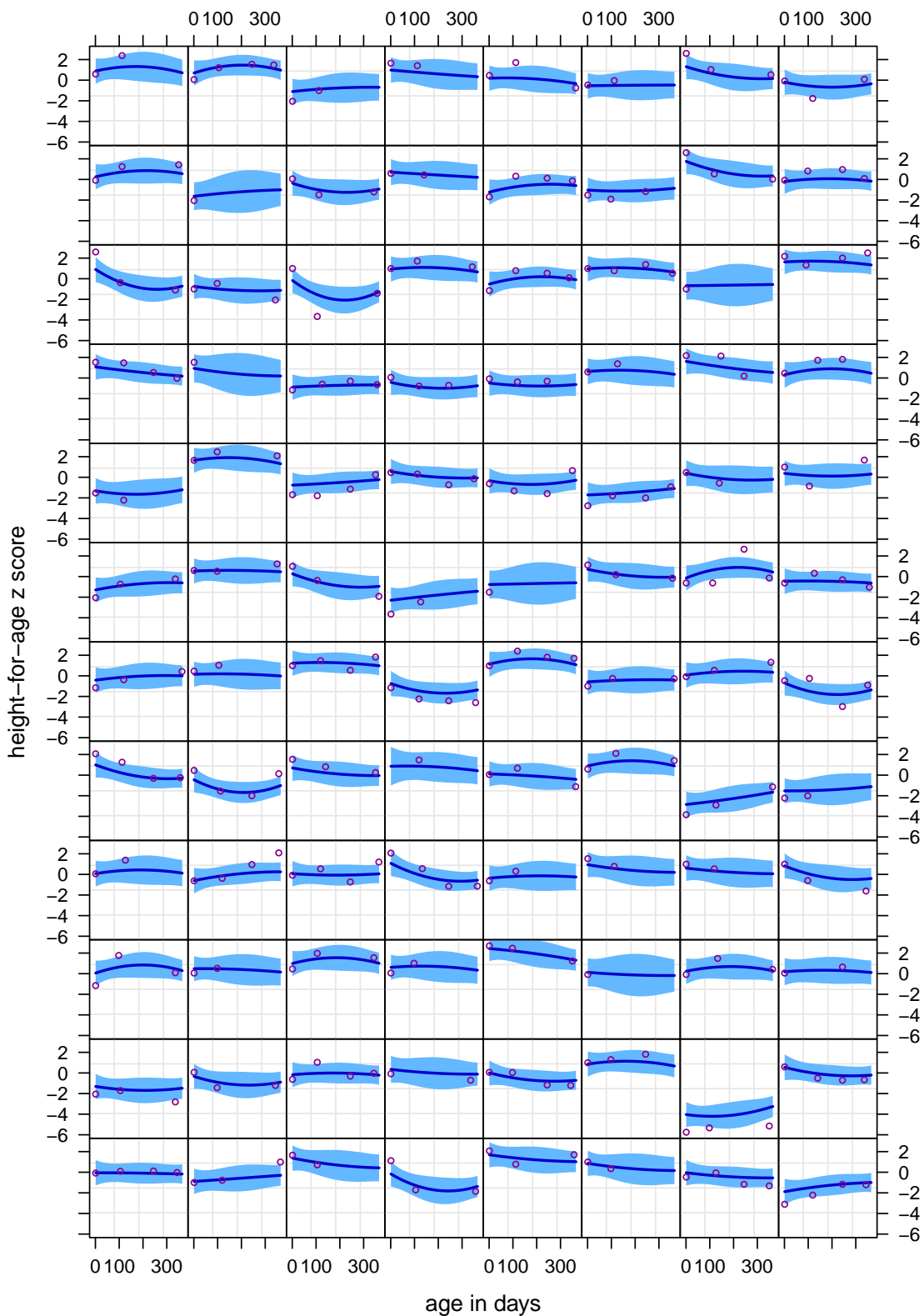$$

(41)

Figure 7: *Fitted random quadratics for 96 randomly chosen infants from the streamlined mean field variational Bayes analysis of data from the Collaborative Perinatal Project for infants in the first year of life. The curves correspond to slices of the fitted surface according to the model defined by (40) and (41) with each of the other predictors set to its average value. The light blue shading corresponds to pointwise 95% credible intervals.*

where y_{ij} denotes the jth response recording for the ith infant and a similar notation applies to the predictors $x_{1ij}, \ldots, x_{7ij}$. The response and predictor variables are:

$y \equiv$ height-for-age z-score (see below for details),

$x_1 \equiv$ age of infant in days,

$x_2 \equiv$ indicator that infant is male,

$x_3 \equiv$ indicator that mother is Asian,

$x_4 \equiv$ indicator that mother is Black,

$x_5 \equiv$ indicator that mother is married,

$x_6 \equiv$ indicator that mother smoked 10 or more cigarettes per day

and $x_7 \equiv$ indicator that mother attended 10 or more ante-natal visits during pregnancy.

The height-for-age z-score is a World Health Organization standardized measure for the height of children after accounting for age. In the Bayesian analysis involving fitting (40) with priors (41) we divided the y and x_1 data by the respective sample standard deviations for each variable. We then convert to the original units for the reporting of results.

Model 40 is an extension of the common random intercepts and slopes model to quadratic fitting, and allows each infant to have his or her own parabola for the effect of age on height-for-age z-score. Figure 7 shows the fits for 96 randomly chosen infants. It is apparant from Figure 7 that the curvature in the age effects warrants the extension to random quadratics.

In Figure 8 we summarize the approximate Bayesian inference for $\beta_3, \ldots, \beta_7$ via 95% credible intervals. The results for Markov chain Monte Carlo-based analysis using the R language (R Core Team, 2018) package rstan (Stan Development Team, 2018) are also shown. The number of mean field variational Bayes iterations is 100 and the Markov chain Monte Carlo results are based on a warmup sample of size $1,000$ and a retained sample of size $1,000$.

It is apparent from Figure 8 that streamlined mean field variational Bayes and Markov chain Monte Carlo deliver very similar inference for the effects of the binary predictors. There are pronounced negative effects due to ethnicity and maternal smoking and a pronounced positive effect due to pre-natal care. Even though the approximate inference is similiar there is a huge difference in computational time. The analyses were run on a `MacBook Air` laptop with a 2.2 gigahertz processor and 8 gigabytes of random access memory. The Markov chain Monte Carlo fits required about 36 hours whilst the streamlined variational results took just 24 seconds.

8 Closing Remarks

We have provided comprehensive coverage of streamlined mean field variational Bayes and variational message passing for two-level and three-level Gaussian response linear mixed models. There are numerous extensions which cannot fit into a single article. One is the addition of penalized spline terms as treated in Lee & Wand (2016). Another is non-Gaussian likelihood fragments. Group specific curve models (e.g. Durban *et al.*, 2005) also lend themselves to streamlining via the SOLVETWOLEVELSPARSELEASTSQUARES and SOLVETHREELEVELSPARSELEASTSQUARES algorithms and Menictas *et al.* (2019) provide full details. Lastly, there are Gaussian response linear mixed models with more than two levels of nesting. The present article provides a blueprint for which these various extensions can be resolved systematically.
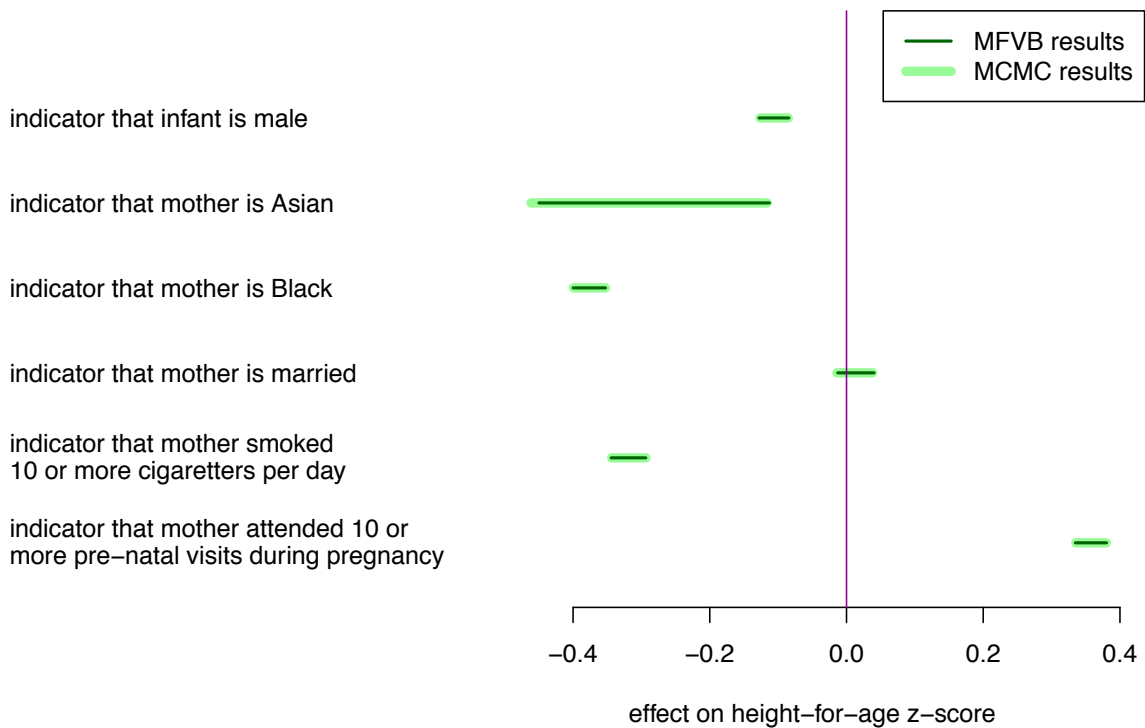
Figure 8: *Approximate 95% credible intervals for $\beta_3, \ldots, \beta_7$ for two approximate Bayesian inference fits to the model defined by (40) and (41) for the data from the Collaborative Perinatal Project for infants in the first year of life. The thin dark green line segments display credible intervals based on streamlined mean field variational Bayes. The thick light green line segments display credible intervals based on a version of Markov chain Monte Carlo.*

Acknowledgments

References

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

Baltagi, B.H. (2013). *Econometric Analysis of Panel Data, Fifth Edition*. Chichester, U.K.: John Wiley & Sons.

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

Durban, M., Harezlak, J., Wand, M.P. and Carroll, R.J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153–1167.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (Editors) (2008). *Longitudinal Data Analysis*. Boca Raton, Florida: Chapman & Hall/CRC.

Gentle, J.E. (2007). *Matrix Algebra*. New York: Springer.

Goldstein, H. (2010). *Multilevel Statistical Models, Fourth Edition*. Chichester, U.K.: John Wiley & Sons.

Harville, D.A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer.

Klebanoff, M.A. (2009). The Collaborative Perinatal Project: a 50-year retrospective. *Paediatric and Perinatal Epidemiology*, **23(1)**, 2–8.

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

McCulloch, C.E., Searle, S.R. & Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models*. Hoboken, New Jersey: John Wiley & Sons.

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. (2005). Divergence measures and message passing. *Microsoft Research Technical Report Series*, **MSR-TR-2005-173**, 1–17.

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

Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-Effects Models in* S *and* S-PLUS. New York: Springer-Verlag.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2. `https://mc-stan.org/`.

Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation, Second Edition*. Hoboken, New Jersey: John Wiley & Sons.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.

Rue, H. & Held, L. (2005). *Gaussian Markov Random Fields*. Boca Raton, Florida: Chapman & Hall/CRC.

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.

Wand, M.P. and Ormerod, J.T. (2011). Penalized wavelets: embedding wavelets into semiparametric regression. *Electronic Journal of Statistics*, **5**, 1654–1717.

A Derivations

A.1 Derivation of Result 1

Straightforward algebra can be used to verify that

$$C^T R_{\text{\tiny BLUP}}^{-1} C + D_{\text{\tiny BLUP}} = B^T B \quad \text{and} \quad C^T R_{\text{\tiny BLUP}}^{-1} C = B^T b$$

where B and b have the sparse forms (9) with non-zero sub-blocks equal to

$$b_i \equiv \begin{bmatrix} \sigma^{-1} y_i \\ 0 \end{bmatrix}, \quad B_i \equiv \begin{bmatrix} \sigma^{-1} X_i \\ O \end{bmatrix}, \quad \text{and} \quad \dot{B}_i \equiv \begin{bmatrix} \sigma^{-1} Z_i \\ \Sigma^{-1/2} \end{bmatrix}.$$

Therefore, in view of (19),

$$\begin{bmatrix} \widehat{\beta} \\ \widehat{u} \end{bmatrix} = (B^T B)^{-1} B^T b = A^{-1} a \quad \text{and} \quad \mathrm{Cov}\left(\begin{bmatrix} \widehat{\beta} \\ \widehat{u} - u \end{bmatrix} \right) = (B^T B)^{-1} = A^{-1}$$

where $A = B^T B$ and $a = B^T b$. Result 1 then follows from extraction of the sub-blocks of $x = A^{-1} a$ and the important sub-blocks of A^{-1} according to (20).

A.2 Derivation of Algorithm 5

Algorithm 5 is simply a proceduralization of Result 1.

A.3 Derivation of Result 2

It is straightforward to verify that the $\mu_{\mathsf{q}(\beta,u)}$ and $\Sigma_{\mathsf{q}(\beta,u)}$ updates, given at (24), may be written as

$$\mu_{\mathsf{q}(\beta,u)} \longleftarrow (B^T B)^{-1} B^T b = A^{-1} a \quad \text{and} \quad \Sigma_{\mathsf{q}(\beta,u)} \longleftarrow (B^T B)^{-1} = A^{-1}$$

where B and b have the forms (9) with

$$b_i \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} y_i \\ m^{-1/2} \Sigma_{\beta}^{-1/2} \mu_{\beta} \\ 0 \end{bmatrix}, \quad B_i \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} X_i \\ m^{-1/2} \Sigma_{\beta}^{-1/2} \\ O \end{bmatrix} \quad \text{and} \quad \dot{B}_i \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} Z_i \\ O \\ M_{\mathsf{q}(\Sigma^{-1})}^{1/2} \end{bmatrix}.$$

A.4 Derivation of Algorithm 6

We first provide expressions for the q-densities for mean field variational Bayesian inference for the parameters in (21), with product density restriction (22). Arguments analogous to those given in, for example, Appendix C of Wand & Ormerod (2011) lead to:

$$\mathsf{q}(\beta, u) \text{ is a } N(\mu_{\mathsf{q}(\beta,u)}, \Sigma_{\mathsf{q}(\beta,u)}) \text{ density function}$$

where

$$\Sigma_{\mathsf{q}(\beta,u)} = (C^T R_{\text{\tiny MFVB}}^{-1} C + D_{\text{\tiny MFVB}})^{-1} \quad \text{and} \quad \mu_{\mathsf{q}(\beta,u)} = \Sigma_{\mathsf{q}(\beta,u)} (C^T R_{\text{\tiny MFVB}}^{-1} y + o_{\text{\tiny MFVB}})$$

with $R_{\text{\tiny MFVB}}$, $D_{\text{\tiny MFVB}}$ and $o_{\text{\tiny MFVB}}$ defined via (25),

$$\mathsf{q}(\sigma^2) \text{ is an Inverse-}\chi^2\left(\xi_{\mathsf{q}(\sigma^2)}, \lambda_{\mathsf{q}(\sigma^2)}\right) \text{ density function}$$

where $\xi_{q(\sigma^2)} = \nu_{\sigma^2} + \sum_{i=1}^m n_i$ and

$$
\begin{aligned}
\lambda_{q(\sigma^2)} &= \mu_{q(1/a_{\sigma^2})} + \sum_{i=1}^m E_q\{\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i\|^2\} \\
&= \mu_{q(1/a_{\sigma^2})} + \sum_{i=1}^m \left[\| E_q(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i)\|^2 + \mathrm{tr}\{\mathrm{Cov}_q(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i)\}\right] \\
&= \mu_{q(1/a_{\sigma^2})} + \sum_{i=1}^m \Big(\| E_q(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i)\|^2 + \mathrm{tr}(\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) + \mathrm{tr}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}) \\
&\qquad\qquad\qquad + 2\,\mathrm{tr}\big[\boldsymbol{Z}_i^T\boldsymbol{X}_i E_q\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{q(\boldsymbol{u}_i)})^T\}\big]\Big)
\end{aligned}
$$

with reciprocal moment $\mu_{q(1/\sigma^2)} = \xi_{q(\sigma^2)}/\lambda_{q(\sigma^2)}$,

$$q(\boldsymbol{\Sigma}) \text{ is an Inverse-G-Wishart} \left(G_{\mathrm{full}}, \xi_{q(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}\right) \text{ density function}$$

where $\xi_{q(\boldsymbol{\Sigma})} = \nu_{\boldsymbol{\Sigma}} + 2q - 2 + m$ and

$$\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})} = \boldsymbol{M}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})} + \sum_{i=1}^m \left(\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}\boldsymbol{\mu}_{q(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{u}_i)}\right)$$

with inverse moment $\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})} = (\xi_{q(\boldsymbol{\Sigma})} - q + 1)\boldsymbol{\Lambda}_{q(\boldsymbol{\Sigma})}^{-1}$,

$$q(a_{\sigma^2}) \text{ is an Inverse-}\chi^2(\xi_{q(a_{\sigma^2})}, \lambda_{q(a_{\sigma^2})}) \text{ density function}$$

where $\xi_{q(a_{\sigma^2})} = \nu_{\sigma^2} + 1$,

$$\lambda_{q(a_{\sigma^2})} = \mu_{q(1/\sigma^2)} + 1/(\nu_{\sigma^2} s_{\sigma^2}^2)$$

with reciprocal moment $\mu_{q(1/a_{\sigma^2})} = \xi_{q(a_{\sigma^2})}/\lambda_{q(a_{\sigma^2})}$ and

$$q(\boldsymbol{A}_{\boldsymbol{\Sigma}}) \text{ is an Inverse-G-Wishart} \left(G_{\mathrm{diag}}, \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}, \boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\right) \text{ density function}$$

where $\xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} = \nu_{\boldsymbol{\Sigma}} + q$,

$$\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})} = \mathrm{diag}\big\{\mathrm{diagonal}\big(\boldsymbol{M}_{q(\boldsymbol{\Sigma}^{-1})}\big)\big\} + \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}$$

with inverse moment $\boldsymbol{M}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})} = \xi_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{\Lambda}_{q(\boldsymbol{A}_{\boldsymbol{\Sigma}})}^{-1}$.

The q-density parameters are interdependent and their Kullback-Leibler divergence optimal values can be found via a coordinate ascent iterative algorithm, which corresponds to Algorithm 2 of Lee & Wand (2016) for the special case of $L = 0$ in the notation used there. However, as explained there, naïve updating of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ has massive computational and storage costs when the number of groups is large. Result 2 asserts that we can instead use SOLVETWOLEVELSPARSELEASTSQUARES (Algorithm 2) to obtain $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and relevant sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$.

A.5 The Approximate Marginal Log-Likelihood for the Two-Level Case

First note that

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

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

$$E_q[\log\{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)\}] = -\tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{m} n_i E_q\{\log(\sigma^2)\}$$

$$-\tfrac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2)}\sum_{i=1}^{m}\left(\|E_{\mathfrak{q}}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i)\|^2 + \mathrm{tr}(\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})})\right.$$

$$+\mathrm{tr}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)})$$

$$\left.+2\,\mathrm{tr}\left[\boldsymbol{Z}_i^T\boldsymbol{X}_i E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}\right]\right).$$

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

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

$$-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right\}\right)$$

$$-\frac{m}{2}E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}|\} - \tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)\right\}\right).$$

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

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

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{p}(\sigma^2|a_{\sigma^2})\}] = E_{\mathfrak{q}}\left(\log\left[\frac{\{1/(2a_{\sigma^2})\}^{\nu_{\sigma^2}/2}}{\Gamma(\nu_{\sigma^2}/2)}(\sigma^2)^{-(\nu_{\sigma^2}/2)-1}\exp\{-1/(2a_{\sigma^2}\sigma^2)\}\right]\right)$$

$$= -\tfrac{1}{2}\nu_{\sigma^2}E_{\mathfrak{q}}\{\log(2a_{\sigma^2})\} - \log\{\Gamma(\tfrac{1}{2}\nu_{\sigma^2})\} - (\tfrac{1}{2}\nu_{\sigma^2} + 1)E_{\mathfrak{q}}\{\log(\sigma^2)\}$$

$$-\tfrac{1}{2}\mu_{\mathfrak{q}(1/a_{\sigma^2})}\mu_{\mathfrak{q}(1/\sigma^2)}.$$

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{q}(\sigma^2)\}] = E_{\mathfrak{q}}\left(\log\left[\frac{\{\lambda_{\mathfrak{q}(\sigma^2)}/2\}^{\xi_{\mathfrak{q}(\sigma^2)}/2}}{\Gamma(\xi_{\mathfrak{q}(\sigma^2)}/2)}(\sigma^2)^{-(\xi_{\mathfrak{q}(\sigma^2)}/2)-1}\exp\{-\lambda_{\mathfrak{q}(\sigma^2)}/(2\sigma^2)\}\right]\right)$$

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

$$-\tfrac{1}{2}\lambda_{\mathfrak{q}(\sigma^2)}\mu_{\mathfrak{q}(1/\sigma^2)}.$$

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{p}(a_{\sigma^2})\}] = E_{\mathfrak{q}}\left(\log\left[\frac{\{1/(2\nu_{\sigma^2}s_{\sigma^2}^2)\}^{1/2}}{\Gamma(1/2)}a_{\sigma^2}^{-(1/2)-1}\exp\{-1/(2\nu_{\sigma^2}s_{\sigma^2}^2 a_{\sigma^2})\}\right]\right)$$

$$= -\tfrac{1}{2}\log(2\nu_{\sigma^2}s_{\sigma^2}^2) - \log\{\Gamma(\tfrac{1}{2})\} - (\tfrac{1}{2} + 1)E_{\mathfrak{q}}\{\log(a_{\sigma^2})\} - \{1/(2\nu_{\sigma^2}s_{\sigma^2}^2)\}\mu_{\mathfrak{q}(1/a_{\sigma^2})}.$$

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{q}(a_{\sigma^2})\}] = E_{\mathfrak{q}}\left(\log\left[\frac{\{\lambda_{\mathfrak{q}(a_{\sigma^2})}/2\}^{\xi_{\mathfrak{q}(a_{\sigma^2})}/2}}{\Gamma(\xi_{\mathfrak{q}(a_{\sigma^2})}/2)}(a_{\sigma^2})^{-(\xi_{\mathfrak{q}(a_{\sigma^2})}/2)-1}\exp\{-\lambda_{\mathfrak{q}(a_{\sigma^2})}/(2a_{\sigma^2})\}\right]\right)$$

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

$$-\tfrac{1}{2}\lambda_{\mathfrak{q}(a_{\sigma^2})}\mu_{\mathfrak{q}(1/a_{\sigma^2})}.$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\log \underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q}) = \sum_{i=1}^{11} T_i$$

where

$$
\begin{aligned}
T_1 \;=\; & -\tfrac{1}{2}\sum_{i=1}^{m} n_i \log(2\pi) \\
& -\tfrac{1}{2}\mu_{\mathsf{q}(1/\sigma^2)}\sum_{i=1}^{m}\Big(\|\,E_{\mathsf{q}}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{u}_i)\|^2+\operatorname{tr}(\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}) \\
& \qquad\qquad +\operatorname{tr}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)})+2\operatorname{tr}\big[\boldsymbol{Z}_i^T\boldsymbol{X}_iE_{\mathsf{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\}\big]\Big), \\[4pt]
T_2 \;=\; & -\tfrac{1}{2}(p+mq)\log(2\pi)-\tfrac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| \\
& -\tfrac{1}{2}\operatorname{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) \\
& -\tfrac{1}{2}\operatorname{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), \\[4pt]
T_3 \;=\; & \tfrac{1}{2}(p+mq)+\tfrac{1}{2}(p+mq)\log(2\pi)+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}|, \\[4pt]
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)}, \\[4pt]
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)}, \\[4pt]
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})} \\[4pt]
T_7 \;=\; & -\tfrac{1}{2}\xi_{\mathsf{q}(a_{\sigma^2})}\log(\lambda_{\mathsf{q}(a_{\sigma^2})}/2)+\log\{\Gamma(\tfrac{1}{2}\xi_{\mathsf{q}(a_{\sigma^2})})\}+\tfrac{1}{2}\lambda_{\mathsf{q}(a_{\sigma^2})}\mu_{\mathsf{q}(1/a_{\sigma^2})}, \\[4pt]
T_8 \;=\; & -\tfrac{1}{2}\operatorname{tr}(\boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\boldsymbol{M}_{\mathsf{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)), \\[4pt]
T_9 \;=\; & -\tfrac{1}{2}(\xi_{\mathsf{q}(\boldsymbol{\Sigma})}-q+1)\log|\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma})}|+\tfrac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma})}\boldsymbol{M}_{\mathsf{q}(\boldsymbol{\Sigma}^{-1})})+\frac{q}{2}(\xi_{\mathsf{q}(\boldsymbol{\Sigma})}+1)\log(2), \\
& +\frac{q}{4}(q-1)\log(\pi)+\sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathsf{q}(\boldsymbol{\Sigma})}+2-j)),
\end{aligned}
$$

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

$$-q\log(2) - \tfrac{q}{4}(q-1)\log(\pi) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(3-j))$$

and
$$T_{11} = -\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} - q + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}| + \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})})$$

$$+\tfrac{q}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} + 1)\log(2) + \tfrac{q}{4}(q-1)\log(\pi) + \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} + 2 - j)).$$

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

$$\mathrm{const} \equiv -\tfrac{1}{2}\sum_{i=1}^{n}n_i\log(2\pi) - \tfrac{1}{2}(p + mq)\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \tfrac{1}{2}(p + mq)$$

$$+\tfrac{1}{2}(p + mq)\log(2\pi) - \tfrac{1}{2}\nu_{\sigma^2}\log(2) + \tfrac{1}{2}(\xi_{\mathfrak{q}(\sigma^2)})\log(2) - \tfrac{1}{2}\log(2\nu_{\sigma^2}s_{\sigma^2}^2) - \log\Gamma(\tfrac{1}{2})$$

$$-\tfrac{1}{2}q(\nu_{\boldsymbol{\Sigma}} + 2q - 1)\log(2) - \tfrac{q}{4}(q-1)\log(\pi) + \tfrac{1}{2}q(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} + 1)\log(2) + \tfrac{q}{4}(q-1)\log(\pi)$$

$$-\tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) - \tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s_{\boldsymbol{\Sigma},j}^2) - q\log(2) - \tfrac{q}{4}(q-1)\log(\pi)$$

$$+\tfrac{1}{2}q(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})} + 1)\log(2) + \tfrac{q}{4}(q-1)\log(\pi)$$

$$= -\tfrac{1}{2}\left(\sum_{i=1}^{m}n_i + 1\right)\log(\pi) - \tfrac{1}{2}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \tfrac{1}{2}(p + mq) + \tfrac{1}{2}(m - 1 + q(\nu_{\boldsymbol{\Sigma}} + q - 1))\log(2)$$

$$-\tfrac{1}{2}\log(\nu_{\sigma^2}) - \tfrac{1}{2}\log(s_{\sigma^2}^2) - \tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) - \tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s_{\sigma^2}^2).$$

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

$$\log\underline{p}(\boldsymbol{y};\mathfrak{q}) = -\tfrac{1}{2}\left(\sum_{i=1}^{m}n_i + 1\right)\log(\pi) - \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \tfrac{1}{2}(p + mq) + \tfrac{1}{2}(m - 1 + q(\nu_{\boldsymbol{\Sigma}} + q - 1))\log(2)$$

$$-\tfrac{1}{2}\log(\nu_{\sigma^2}) - \tfrac{1}{2}\log(s_{\sigma^2}^2) - \tfrac{1}{2}q(2-q)\log(\nu_{\boldsymbol{\Sigma}}) - \tfrac{1}{2}(2-q)\sum_{j=1}^{q}\log(s_{\sigma^2}^2) - \log\{\Gamma(\tfrac{1}{2}\nu_{\sigma^2})\}$$

$$-\tfrac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2)}\sum_{i=1}^{m}\left(\|E_{\mathfrak{q}}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{u}_i)\|^2 + \mathrm{tr}(\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}) + \mathrm{tr}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)})\right.$$

$$\left.+2\,\mathrm{tr}\left[\boldsymbol{Z}_i^T\boldsymbol{X}_iE_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}\right]\right) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} + 2 - j))$$

$$-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right\}\right) + \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} - q + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}|$$

$$-\tfrac{1}{2}\mathrm{tr}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\right)\right\}\right) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}| - \tfrac{1}{2}\mu_{\mathfrak{q}(1/a_{\sigma^2})}\mu_{\mathfrak{q}(1/\sigma^2)}$$

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

$$-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}) - \sum_{j=1}^{q}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}} + 2q - j)) - \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})})$$

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

From Theorem 1 of Nolan & Wand (2018), the $\log |\Sigma_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}|$ term has the following stream-lined form:

$$\log |\Sigma_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}| = \sum_{i=1}^{m} \log |\Sigma_{\mathsf{q}(\boldsymbol{u}_i)}|$$
$$+ \log \left| \Sigma_{\mathsf{q}(\boldsymbol{\beta})} - \sum_{i=1}^{m} E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\} \Sigma_{\mathsf{q}(\boldsymbol{u}_i)}^{-1} E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\}^T \right|.$$

A.6 Derivation of Result 3

Note that

$$\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u}) \propto m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \, m_{(\boldsymbol{\beta},\boldsymbol{u}) \to \mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u})$$

$$= \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1 \leq i \leq m}{\mathrm{stack}} \begin{bmatrix} \boldsymbol{u}_i \\ \mathrm{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})} \right\}$$

$$= \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{a} - \frac{1}{2} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{A} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right\}$$

where $\boldsymbol{a}$ and $\boldsymbol{A}$ as given in Result 3 and the last step uses facts such as $\mathrm{vech}(\boldsymbol{M}) = \boldsymbol{D}_d^+ \mathrm{vec}(\boldsymbol{M})$ for any symmetric $d \times d$ matrix $\boldsymbol{M}$. Standard manipulations then lead to

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}\boldsymbol{a} \quad \text{and} \quad \Sigma_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}.$$

Result 3 then follows from extraction of the sub-blocks of $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{a}$ and the important sub-blocks of $\boldsymbol{A}^{-1}$ according to (26).

A.7 Derivation of Algorithm 7

The two-level reduced exponential family form is

$$\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u}) \propto \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1 \leq i \leq m}{\mathrm{stack}} \begin{bmatrix} \boldsymbol{u}_i \\ \mathrm{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} \right\}$$

$$= \exp \left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{a} - \frac{1}{2} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{A} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right\}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are as defined in Result 3 with $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow (\boldsymbol{\beta},\boldsymbol{u})}$ replaced by $\boldsymbol{\eta}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with $\boldsymbol{A}$ having two-level sparse structure. As with the derivation of Result 3, we have the relationships

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}\boldsymbol{a} \quad \text{and} \quad \Sigma_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}. \tag{42}$$

The first part of Algorithm 7 is such that the entries of $\boldsymbol{\eta}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ are sequentially unpacked and stored in the vectors $\boldsymbol{\omega}_{12}$ and $\boldsymbol{\omega}_{15i}$, $1 \leq i \leq m$, corresponding to the $\boldsymbol{a}$ vector according

50

to the partitioning in (7) and the matrices $\mathbf{\Omega}_{14}$ and $\mathbf{\Omega}_{18i}, \mathbf{\Omega}_{19i}, 1 \leq i \leq m$, corresponding to the non-zero sub-blocks of $\mathbf{A}$ in (7).

Next, $\mathcal{S}_3$ stores the streamlined solution to (42) according to the SOLVETWOLEVELSPARSE-MATRIX algorithm (Algorithm 1). The remainder of Algorithm 7 is plucking off the relevant common parameter sub-blocks of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ based (42) and keeping in mind that (42) represents a two-level sparse matrix problem.

A.8 Derivation of Algorithm 8

First note that the logarithm of the fragment factor is, as a function of $(\boldsymbol{\beta}, \boldsymbol{u})$:

$$\log \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} \|\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \boldsymbol{Z}_i \boldsymbol{u}_i\|^2 + \text{const}$$

$$= (1/\sigma^2) \begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \text{stack}_{1\leq i\leq m} \begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \begin{bmatrix} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{y}_i \\ -\frac{1}{2}\sum_{i=1}^{m} \boldsymbol{D}_p^T \text{vec}(\boldsymbol{X}_i^T \boldsymbol{X}_i) \\ \text{stack}_{1\leq i\leq m} \begin{bmatrix} \boldsymbol{Z}_i^T \boldsymbol{y}_i \\ -\frac{1}{2}\boldsymbol{D}_q^T \text{vec}(\boldsymbol{Z}_i^T \boldsymbol{Z}_i) \\ -\text{vec}(\boldsymbol{X}_i^T \boldsymbol{Z}_i) \end{bmatrix} \end{bmatrix} + \text{const.}$$

Therefore, from equations (8) and (9) of Wand (2017),

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \longleftarrow \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \text{stack}_{1\leq i\leq m} \begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})} \equiv \mu_{\mathfrak{q}(1/\sigma^2)} \begin{bmatrix} \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{y}_i \\ -\frac{1}{2}\sum_{i=1}^{m} \boldsymbol{D}_p^T \text{vec}(\boldsymbol{X}_i^T \boldsymbol{X}_i) \\ \text{stack}_{1\leq i\leq m} \begin{bmatrix} \boldsymbol{Z}_i^T \boldsymbol{y}_i \\ -\frac{1}{2}\boldsymbol{D}_q^T \text{vec}(\boldsymbol{Z}_i^T \boldsymbol{Z}_i) \\ -\text{vec}(\boldsymbol{X}_i^T \boldsymbol{Z}_i) \end{bmatrix} \end{bmatrix}$$

and $\mu_{\mathfrak{q}(1/\sigma^2)}$ denotes expectation of $1/\sigma^2$ with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to \sigma^2}(\sigma^2) \, m_{\sigma^2 \to \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\sigma^2)$$

which is an Inverse χ^2 density function with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2}$ and, according to Table S.1 in the online supplement of Wand (2017), leads to

$$\mu_{\mathfrak{q}(1/\sigma^2)} \longleftarrow \left((\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2})_1 + 1 \right) / (\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \leftrightarrow \sigma^2})_2.$$

The other factor to stochastic node message update is

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to\sigma^2}(\sigma^2) \longleftarrow \exp\left\{\begin{bmatrix}\log(\sigma^2)\\[4pt]1/\sigma^2\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to\sigma^2}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to\sigma^2} \equiv \begin{bmatrix}-\frac{1}{2}\displaystyle\sum_{i=1}^m n_i\\[12pt]-\frac{1}{2}\displaystyle\sum_{i=1}^m E_{\mathsf{q}}\{\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{u}_i\|^2\}\end{bmatrix}$$

with E_{q} denoting expectation with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\, m_{(\boldsymbol{\beta},\boldsymbol{u})\to\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u}).$$

Then note that

$$E_{\mathsf{q}}\{\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{u}_i\|^2\} = \|\boldsymbol{y}_i-\boldsymbol{X}_i\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_i\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)}\|^2 + \operatorname{tr}(\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})})$$
$$+\operatorname{tr}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)}) + 2\operatorname{tr}\big[\boldsymbol{Z}_i^T\boldsymbol{X}_i E_{\mathsf{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i)})^T\}\big]$$

where, for example, $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} \equiv E_{\mathsf{q}}(\boldsymbol{\beta})$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i)} \equiv \operatorname{Cov}_{\mathsf{q}}(\boldsymbol{u}_i)$. Result 3 links sub-blocks of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})}$ with the required sub-vectors of $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and sub-blocks of $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$. These matrices are extracted from $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})}$ in the call to TwoLevelNaturalTo-CommonParameters algorithm (Algorithm 7).

A.9 Derivation of Result 4

The derivation of Result 4 is very similar to that for Result 3.

A.10 Derivation of Algorithm 9

The logarithm on the fragment factor is, as a function of $(\boldsymbol{\beta},\boldsymbol{u})$:

$$\log\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}) = -\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}) - \frac{1}{2}\sum_{i=1}^m \boldsymbol{u}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_i + \text{const}$$

$$= \begin{bmatrix}\boldsymbol{\beta}\\[4pt]\operatorname{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\[4pt]\underset{1\le i\le m}{\operatorname{stack}}\begin{bmatrix}\boldsymbol{u}_i\\[4pt]\operatorname{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\\[4pt]\operatorname{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T)\end{bmatrix}\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\\[4pt]-\frac{1}{2}\boldsymbol{D}_p^T\operatorname{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})\\[4pt]\underset{1\le i\le m}{\operatorname{stack}}\begin{bmatrix}\boldsymbol{0}_q\\[4pt]-\frac{1}{2}\boldsymbol{D}_q^T\operatorname{vec}(\boldsymbol{\Sigma}^{-1})\\[4pt]\boldsymbol{0}_{pq}\end{bmatrix}\end{bmatrix} + \text{const}.$$

Therefore, from equations (8) and (9) of Wand (2017),

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \longleftarrow \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\[4pt]\operatorname{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\[4pt]\underset{1\le i\le m}{\operatorname{stack}}\begin{bmatrix}\boldsymbol{u}_i\\[4pt]\operatorname{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\\[4pt]\operatorname{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T)\end{bmatrix}\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to(\boldsymbol{\beta},\boldsymbol{u})}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to(\boldsymbol{\beta},\boldsymbol{u})} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\frac{1}{2}\boldsymbol{D}_p^T\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\ \underset{1\le i\le m}{\mathrm{stack}} \begin{bmatrix} \boldsymbol{0}_q \\ -\frac{1}{2}\boldsymbol{D}_q^T\mathrm{vec}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}) \\ \boldsymbol{0}_{pq} \end{bmatrix} \end{bmatrix}$$

and $\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}$ denotes expectation of $\boldsymbol{\Sigma}^{-1}$ with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma})\,m_{\boldsymbol{\Sigma}\to\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{\Sigma})$$

which is an Inverse G-Wishart density function with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\leftrightarrow\boldsymbol{\Sigma}}$ and, according to Table S.1 in the online supplement of Wand (2017), leads to

$$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} \longleftarrow \{\omega_{23}+\tfrac{1}{2}(q+1)\}\{\mathrm{vec}^{-1}(\boldsymbol{\omega}_{24})\}^{-1}$$

where ω_{23} is the first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\leftrightarrow\boldsymbol{\Sigma}}$ and $\boldsymbol{\omega}_{24}$ is the vector containing the remaining entries of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\leftrightarrow\boldsymbol{\Sigma}}$.

The other factor to stochastic node message update is

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}) \longleftarrow \exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}| \\ \mathrm{vech}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}} \equiv \begin{bmatrix} -\frac{1}{2}m \\ -\frac{1}{2}\sum_{i=1}^{m}\boldsymbol{D}_q^T\mathrm{vec}\{E_{\mathfrak{q}}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\} \end{bmatrix}$$

with $E_{\mathfrak{q}}$ denoting expectation with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\,m_{(\boldsymbol{\beta},\boldsymbol{u})\to\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})}(\boldsymbol{\beta},\boldsymbol{u}).$$

Then note that
$$E_{\mathfrak{q}}(\boldsymbol{u}_i\boldsymbol{u}_i^T) = \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}$$

where, as before, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)} \equiv E_{\mathfrak{q}}(\boldsymbol{u}_i)$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \equiv \mathrm{Cov}_{\mathfrak{q}}(\boldsymbol{u}_i)$. Result 4 links sub-blocks of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to(\boldsymbol{\beta},\boldsymbol{u})}$ with the required sub-vectors of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$. We then call upon Algorithm 7 to obtain $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}$, $1 \le i \le m$.

A.11   Derivation of Result 5

If $\boldsymbol{B}$ and $\boldsymbol{b}$ have the forms given by (15) with

$$\boldsymbol{b}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{y}_{ij} \\ 0 \\ 0 \end{bmatrix},\ \boldsymbol{B}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{X}_{ij} \\ \boldsymbol{O} \\ \boldsymbol{O} \end{bmatrix},\ \dot{\boldsymbol{B}}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_{ij}^{\mathrm{L1}} \\ n_i^{-1/2}(\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1/2} \\ \boldsymbol{O} \end{bmatrix}\ \text{and}\ \ddot{\boldsymbol{B}}_{ij} \equiv \begin{bmatrix} \sigma^{-1}\boldsymbol{Z}_{ij}^{\mathrm{L2}} \\ \boldsymbol{O} \\ (\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1/2} \end{bmatrix},$$

then straightforward algebra leads to

$$\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{C}^T\boldsymbol{R}_{\mathrm{BLUP}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\mathrm{BLUP}} \quad\text{and}\quad \boldsymbol{B}^T\boldsymbol{b} = \boldsymbol{C}^T\boldsymbol{R}_{\mathrm{BLUP}}^{-1}\boldsymbol{y}$$

where $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{R}_{\mathrm{BLUP}}$ are given by (35). The remainder of the derivation of Result 5 is analogous to that for Result 1.

A.12 Derivation of Algorithm 10

Algorithm 10 is simply a proceduralization of Result 5.

A.13 Derivation of Result 6

Routine matrix algebraic steps can verify that the $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ updates,

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T \boldsymbol{R}_{\text{\tiny MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{\tiny MFVB}})^{-1}(\boldsymbol{C}^T \boldsymbol{R}_{\text{\tiny MFVB}}^{-1} \boldsymbol{y} + \boldsymbol{o}_{\text{\tiny MFVB}}) \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T \boldsymbol{R}_{\text{\tiny MFVB}}^{-1} \boldsymbol{C} + \boldsymbol{D}_{\text{\tiny MFVB}})^{-1},$$

with $\boldsymbol{C}$, $\boldsymbol{D}_{\text{\tiny MFVB}}$ and $\boldsymbol{R}_{\text{\tiny MFVB}}$ as defined at (35), may be written as

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{b} = \boldsymbol{A}^{-1} \boldsymbol{a} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow (\boldsymbol{B}^T \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1}$$

where $\boldsymbol{B}$ and $\boldsymbol{b}$ have the sparse three-level forms given by (15) with

$$\boldsymbol{b}_{ij} \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} \boldsymbol{y}_{ij} \\ \left(\sum_{i=1}^{m} n_i \right)^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{B}_{ij} \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} \boldsymbol{X}_{ij} \\ \left(\sum_{i=1}^{m} n_i \right)^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{O} \\ \boldsymbol{O} \end{bmatrix},$$

$$\dot{\boldsymbol{B}}_{ij} \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} \boldsymbol{Z}_{ij}^{\text{\tiny L1}} \\ \boldsymbol{O} \\ n_i^{-1/2} \left(\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\text{\tiny L1}})^{-1})} \right)^{1/2} \\ \boldsymbol{O} \end{bmatrix} \quad \text{and} \quad \ddot{\boldsymbol{B}}_{ij} \equiv \begin{bmatrix} \mu_{\mathsf{q}(1/\sigma^2)}^{1/2} \boldsymbol{Z}_{ij}^{\text{\tiny L2}} \\ \boldsymbol{O} \\ \boldsymbol{O} \\ \left(\boldsymbol{M}_{\mathsf{q}((\boldsymbol{\Sigma}^{\text{\tiny L2}})^{-1})} \right)^{1/2} \end{bmatrix}.$$

A.14 Derivation of Algorithm 11

Algorithm 11 is the three-level counterpart of Algorithm 6 and its derivation is analogous to that given for Algorithm 6 in Section A.4.

The first difference is that the $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ updates are expressible as three-level sparse matrix least squares problems and so the SolveThreeLevelSparseLeastSquares algorithm (Algorithm 4) is used for streamlined updating of their relevant sub-blocks.

We still have $\mathsf{q}(\sigma^2)$ optimally being an Inverse Chi-Squared density function but with shape parameter

$$\xi_{\mathsf{q}(\sigma^2)} = \nu_{\sigma^2} + \tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} o_{ij}$$

and rate parameter

$$\lambda_{\mathsf{q}(\sigma^2)} = \mu_{\mathsf{q}(1/a_{\sigma^2})} + \tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} E_{\mathsf{q}}\{\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}_{ij}^{\text{\tiny L1}}\boldsymbol{u}_i^{\text{\tiny L1}} - \boldsymbol{Z}_{ij}^{\text{\tiny L2}}\boldsymbol{u}_{ij}^{\text{\tiny L2}}\|^2\}$$

$$= \mu_{\mathsf{q}(1/a_{\sigma^2})} + \tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Big(\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_{ij}^{\text{\tiny L1}}\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{\tiny L1}})} - \boldsymbol{Z}_{ij}^{\text{\tiny L2}}\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{\tiny L2}})}\|^2$$

$$+ \text{tr}(\boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij} \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}) + \text{tr}\{(\boldsymbol{Z}_{ij}^{\text{\tiny L1}})^T \boldsymbol{Z}_{ij}^{\text{\tiny L1}} \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\text{\tiny L1}})}\} + \text{tr}\{(\boldsymbol{Z}_{ij}^{\text{\tiny L2}})^T \boldsymbol{Z}_{ij}^{\text{\tiny L2}} \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{\tiny L2}})}\}$$

$$+ 2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{\tiny L1}})^T \boldsymbol{X}_{ij} E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{\tiny L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{\tiny L1}})})^T\}\big]$$

$$+ 2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{\tiny L2}})^T \boldsymbol{X}_{ij} E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{\tiny L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{\tiny L2}})})^T\}\big]$$

$$+ 2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{\tiny L2}})^T \boldsymbol{Z}_{ij}^{\text{\tiny L1}} E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\text{\tiny L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\text{\tiny L1}})})(\boldsymbol{u}_{ij}^{\text{\tiny L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\text{\tiny L2}})})^T\}\big]\Big).$$

The optimal $\mathsf{q}(a_{\sigma^2})$ density function is unaffected by the change from the two-level case to the three-level situation.

The random effects covariance matrices are such that

$$\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}}) \text{ is an Inverse-G-Wishart}\left(G_{\mathrm{full}}, \xi_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})}, \boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})}\right) \text{ density function}$$

where $\xi_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} = \nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}} + 2q_1 - 2 + m$ and

$$\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} = \boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}^{-1})} + \sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}^{T} + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\right),$$

whilst

$$\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}}) \text{ is an Inverse-G-Wishart}\left(G_{\mathrm{full}}, \xi_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}, \boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}\right) \text{ density function}$$

where $\xi_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} = \nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}} + 2q_2 - 2 + \sum_{i=1}^{m} n_i$ and

$$\boldsymbol{\Lambda}_{\mathsf{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} = \boldsymbol{M}_{\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}^{-1})} + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}^{T} + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\right).$$

The optimal $\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})$ and $\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})$ density functions have the same derivations and forms as $\mathsf{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})$ in the two-level case.

Algorithm 11 is a streamlined iterative coordinate ascent for determination of Kullback-Leibler optimal values of each of the q-density parameters in the Bayesian three-level mixed model (37).

A.15 The Approximate Marginal Log-Likelihood for the Three-Level Case

First note that

$$
\begin{aligned}
\log \underline{p}(\boldsymbol{y}; \mathsf{q}) &= E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{L1}}, \boldsymbol{u}^{\mathrm{L2}}, \sigma^2, a_{\sigma^2}, \boldsymbol{\Sigma}^{\mathrm{L1}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}, \boldsymbol{\Sigma}^{\mathrm{L2}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}) \\
&\qquad - \log \mathsf{q}^{*}(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{L1}}, \boldsymbol{u}^{\mathrm{L2}}, \sigma^2, a_{\sigma^2}, \boldsymbol{\Sigma}^{\mathrm{L1}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}, \boldsymbol{\Sigma}^{\mathrm{L2}}, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})\} \\
&= E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{L1}}, \boldsymbol{u}^{\mathrm{L2}}, \sigma^2)\} + E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{L1}}, \boldsymbol{u}^{\mathrm{L2}} \,|\, \boldsymbol{\Sigma}^{\mathrm{L1}}, \boldsymbol{\Sigma}^{\mathrm{L2}})\} \\
&\quad - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(\boldsymbol{\beta}, \boldsymbol{u}^{\mathrm{L1}}, \boldsymbol{u}^{\mathrm{L2}})\} + E_{\mathsf{q}}\{\log \mathfrak{p}(\sigma^2 \,|\, a_{\sigma^2})\} - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(\sigma^2)\} \\
&\quad + E_{\mathsf{q}}\{\log \mathfrak{p}(a_{\sigma^2})\} - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(a_{\sigma^2})\} + E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{\Sigma}^{\mathrm{L1}} \,|\, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})\} - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(\boldsymbol{\Sigma}^{\mathrm{L1}})\} \\
&\quad + E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})\} - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})\} + E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{\Sigma}^{\mathrm{L2}} \,|\, \boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})\} - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(\boldsymbol{\Sigma}^{\mathrm{L2}})\} \\
&\quad + E_{\mathsf{q}}\{\log \mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})\} - E_{\mathsf{q}}\{\log \mathsf{q}^{*}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})\}.
\end{aligned}
$$

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

$$
\begin{aligned}
E_{\mathsf{q}}[\log\{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)\}] &= -\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} o_{ij} \log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} o_{ij}\, E_{\mathsf{q}}\{\log(\sigma^2)\} \\
&\quad - \tfrac{1}{2}\mu_{\mathsf{q}(1/\sigma^2)}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big(\| E_{\mathsf{q}}(\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{u}_{ij}^{\mathrm{L2}})\|^2 \\
&\qquad + \mathrm{tr}(\boldsymbol{X}_{ij}^{T}\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}) + \mathrm{tr}((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}) + \mathrm{tr}((\boldsymbol{Z}_{ij}^{\mathrm{L2}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}) \\
&\qquad + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{X}_{ij}E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^{T}\}\big] \\
&\qquad + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L2}}E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^{T}\}\big]\Big).
\end{aligned}
$$

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}^{\text{L1}}, \boldsymbol{u}^{\text{L2}} \mid \boldsymbol{\Sigma}^{\text{L1}}, \boldsymbol{\Sigma}^{\text{L2}})\}] = -\tfrac{1}{2}(p + mq_1 + q_2 \sum_{i=1}^{m} n_i) \log(2\pi) - \tfrac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| - \frac{m}{2}E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}^{\text{L1}}|\}$$

$$-\tfrac{1}{2}\sum_{i=1}^{m} n_i E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}^{\text{L2}}|\}$$

$$-\tfrac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^{T} + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right\}\right)$$

$$-\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}^{T} + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\right)\right\}\right)$$

$$-\tfrac{1}{2}\text{tr}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^{T} + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\right)\right\}\right).$$

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}^{\text{L1}}, \boldsymbol{u}^{\text{L2}})\}] = -\tfrac{1}{2}(p + mq_1 + q_2 \sum_{i=1}^{m} n_i) - \tfrac{1}{2}(p + mq_1 + q_2 \sum_{i=1}^{m} n_i)\log(2\pi)$$

$$-\tfrac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}^{\text{L1}}, \boldsymbol{u}^{\text{L2}})}\right|.$$

The fourth to seventh terms in $\log \underline{p}(\boldsymbol{y}; q)$ are as given in the two-level case. The eighth to eleventh terms in $\log \underline{p}(\boldsymbol{y}; q)$ are as given in the two-level case, but, with $\boldsymbol{\Sigma}$, $\boldsymbol{A}_{\boldsymbol{\Sigma}}$, $\nu_{\boldsymbol{\Sigma}}$, $s_{\boldsymbol{\Sigma}, j}$, q replaced with $\boldsymbol{\Sigma}^{\text{L1}}$, $\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}}$, $\nu_{\boldsymbol{\Sigma}^{\text{L1}}}$, $s_{\boldsymbol{\Sigma}^{\text{L1}}, j}$, q_1 respectively. The twelfth of the $\log \underline{p}(\boldsymbol{y}; q)$ terms is

$$E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{\Sigma}^{\text{L2}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})\}] = E_{\mathfrak{q}}\left(\frac{|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}|^{-\frac{1}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + q_2 - 1)}|\boldsymbol{\Sigma}^{\text{L2}}|^{-\frac{1}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2)}}{2^{\frac{q_2}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2 - 1)}\pi^{\frac{q_2}{4}(q_2 - 1)}\prod_{j=1}^{q_2}\Gamma(\frac{1}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2 - j))}\right.$$

$$\left.\times \exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}^{-1}(\boldsymbol{\Sigma}^{\text{L2}})^{-1})\}\right)$$

$$= -\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + q_2 - 1)E_{\mathfrak{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}|\} - \tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2)E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}^{\text{L2}}|\}$$

$$-\tfrac{1}{2}\text{tr}(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}^{-1})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}) - \frac{q_2}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2 - 1)\log(2)$$

$$-\frac{q_2}{4}(q_2 - 1)\log(\pi) - \sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}^{\text{L2}}} + 2q_2 - j)).$$

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

$$E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})\}] = E_{\mathfrak{q}}\left(\frac{|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}|^{\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} - q_2 + 1)}|\boldsymbol{\Sigma}^{\text{L2}}|^{-\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} + 2)}}{2^{\frac{q_2}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} + 1)}\pi^{\frac{q_2}{4}(q_2 - 1)}\prod_{j=1}^{q_2}\Gamma(\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} + 2 - j))}\right.$$

$$\left.\times \exp\{-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}(\boldsymbol{\Sigma}^{\text{L2}})^{-1})\}\right)$$

$$= \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} - q_2 + 1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}| - \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} + 2)E_{\mathfrak{q}}\{\log|\boldsymbol{\Sigma}^{\text{L2}}|\}$$

$$-\tfrac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}) - \frac{q_2}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} + 1)\log(2) - \frac{q_2}{4}(q_2 - 1)\log(\pi)$$

$$-\sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})} + 2 - j)).$$

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

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})\}] \;=\; & E_{\mathfrak{q}}\left(\frac{|\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}}|^{\frac{1}{2}(2-q_2)}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}|^{-3/2}}{2^{q_2}\pi^{\frac{q_2}{4}(q_2-1)}\prod_{j=1}^{q}\Gamma(\frac{1}{2}(3-j))}\exp\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}}\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})\}\right) \\
=\; & -\tfrac{1}{2}q_2(2-q_2)\log(\nu_{\boldsymbol{\Sigma}^{L2}}) - \tfrac{1}{2}(2-q_2)\sum_{j=1}^{q_2}\log(s^2_{\boldsymbol{\Sigma}^{L2},j}) - \frac{3}{2}E_{\mathfrak{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}|\} \\
& -\tfrac{1}{2}\sum_{j=1}^{q_2}1/(\nu_{\boldsymbol{\Sigma}^{L2}}s^2_{\boldsymbol{\Sigma}^{L2},j})\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})}\right)_{jj} - q_2\log(2) - \frac{q_2}{4}(q_2-1)\log(\pi) \\
& -\sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(3-j)).
\end{aligned}
$$

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

$$
\begin{aligned}
E_{\mathfrak{q}}[\log\{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})\}] \;=\; & E_{\mathfrak{q}}\left(\frac{|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}|^{\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}-q_2+1)}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}|^{-\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+2)}}{2^{\frac{q_2}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+1)}\pi^{\frac{q_2}{4}(q_2-1)}\prod_{j=1}^{q_2}\Gamma(\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+2-j))}\right. \\
& \left.\times\exp\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}^{-1}\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})\}\right) \\
=\; & \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}-q_2+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}| - \tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+2)E_{\mathfrak{q}}\{\log|\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}|\} \\
& -\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})}) - \frac{q_2}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+1)\log(2) - \frac{q_2}{4}(q_2-1)\log(\pi) \\
& -\sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+2-j)).
\end{aligned}
$$

Similar cancellations as shown in the two-level case of constant terms are made and the final expression is:

$$\log \underline{\mathfrak{p}}(\boldsymbol{y}; \mathfrak{q}) =$$

$$-\frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i}o_{ij}+1\right)\log(\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \frac{1}{2}(p + mq_1 + q_2\sum_{i=1}^{m}n_i) - \frac{1}{2}\log(\nu_{\sigma^2}s_{\sigma^2}^2)$$

$$-\frac{1}{2}q_1(2-q_1)\log(\nu_{\boldsymbol{\Sigma}^{L1}}) - \frac{1}{2}(2-q_1)\sum_{j=1}^{q_1}\log(s_{\boldsymbol{\Sigma}^{L1},j}) - \frac{1}{2}q_2(2-q_2)\log(\nu_{\boldsymbol{\Sigma}^{L2}}) + \frac{1}{2}\nu_{\sigma^2}\log(2)$$

$$-\frac{1}{2}(2-q_2)\sum_{j=1}^{q_2}\log(s_{\boldsymbol{\Sigma}^{L2},j}) + \frac{1}{2}q_1(m-1+q_1+\nu_{\boldsymbol{\Sigma}^{L1}})\log(2) + \frac{1}{2}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}^{L1},\boldsymbol{u}^{L2})}\right|$$

$$+\frac{q_2}{2}\left(\sum_{i=1}^{m}n_i - 1 + q_2 + \nu_{\boldsymbol{\Sigma}^{L2}}\right)\log(2) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left\{\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}\right)^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\right\}\right)$$

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

$$-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{L1})^{-1})}\left\{\sum_{i=1}^{m}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})}\right)\right\}\right) - \frac{1}{2}\sum_{j=1}^{q_1}\frac{1}{\nu_{\boldsymbol{\Sigma}^{L1}}s_{\boldsymbol{\Sigma}^{L1},j}^2}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}}^{-1})}\right)_{jj}$$

$$-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{L2})^{-1})}\left\{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})}\right)\right\}\right) + \frac{1}{2}\lambda_{\mathfrak{q}(\sigma^2)}\mu_{\mathfrak{q}(1/\sigma^2)}$$

$$-\{1/(2\nu_{\sigma^2}s_{\sigma^2}^2)\}\mu_{\mathfrak{q}(1/a_{\sigma^2})} - \frac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})}\log(\lambda_{\mathfrak{q}(a_{\sigma^2})}) + \log\{\Gamma(\tfrac{1}{2}\xi_{\mathfrak{q}(a_{\sigma^2})})\} + \frac{1}{2}\lambda_{\mathfrak{q}(a_{\sigma^2})}\mu_{\mathfrak{q}(1/a_{\sigma^2})}$$

$$-\sum_{j=1}^{q_1}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}^{L1}}+2q_1-j)) - \frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{L1})}-q_1+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{L1})}| - \sum_{j=1}^{q_1}\log\Gamma(\tfrac{1}{2}(3-j))$$

$$+\sum_{j=1}^{q_1}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{L1})}+2-j)) + \sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}+2-j)) - \sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(3-j))$$

$$-\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}})}-q_1+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}})}| + \sum_{j=1}^{q_1}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}})}+2-j))$$

$$-\sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(\nu_{\boldsymbol{\Sigma}^{L2}}+2q_2-j)) - \frac{1}{2}\mathrm{tr}\left\{\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{L1})^{-1})}\left(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}})^{-1})} - \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{L1})}\right)\right\}$$

$$-\frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{L2})}-q_2+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{L2})}| + \sum_{j=1}^{q_2}\log\Gamma(\tfrac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{L2})}+2-j))$$

$$-\frac{1}{2}\sum_{j=1}^{q_2}\frac{1}{\nu_{\boldsymbol{\Sigma}^{L2}}s_{\boldsymbol{\Sigma}^{L2},j}^2}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})}\right)_{jj} - \frac{1}{2}(\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}-q_2+1)\log|\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}|$$

$$-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{L2})^{-1})}\left(\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})} - \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{L2})}\right)\right\} + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}})}\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L2}}^{-1})})$$

$$+\frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}})}\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{A}_{\boldsymbol{\Sigma}^{L1}})^{-1})})$$

$$-\frac{1}{2}\mu_{\mathfrak{q}(1/\sigma^2)}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\|E_{\mathfrak{q}}(\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}_{ij}^{L1}\boldsymbol{u}_i^{L1} - \boldsymbol{Z}_{ij}^{L2}\boldsymbol{u}_{ij}^{L2})\|^2 + \mathrm{tr}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})})\right.$$

$$+\mathrm{tr}((\boldsymbol{Z}_{ij}^{L1})^T\boldsymbol{Z}_{ij}^{L1}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})}) + \mathrm{tr}((\boldsymbol{Z}_{ij}^{L2})^T\boldsymbol{Z}_{ij}^{L2}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})})$$

$$+2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{L1})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{L1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})})^T\}\big]$$

$$\left. +2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{L1})^T\boldsymbol{Z}_{ij}^{L2}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{L1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})})(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})})^T\}\big]\right).$$

Note that the $\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}^{L1},\boldsymbol{u}^{L2})}\right|$ term involves the determinant of a potentially very large matrix, but with three-level sparse structure. From Theorem 3 of Nolan & Wand (2018), a

streamlined expression for $\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}^{\text{L1}},\boldsymbol{u}^{\text{L2}})}\right|$ is

$$
\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}^{\text{L1}},\boldsymbol{u}^{\text{L2}})}\right|
$$

$$
= \sum_{i=1}^{m}\left[\sum_{j=1}^{n_i}\log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\right| + \log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} - \sum_{j=1}^{n_i}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^{-1}\right.\right.
$$

$$
\left.\left.\times E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}^T\right|\right]
$$

$$
+ \log\left|\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} - \sum_{i=1}^{m}\sum_{j=1}^{n_i}E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^{-1}E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}^T\right.
$$

$$
\left. - \sum_{i=1}^{m}\boldsymbol{H}_{12,i}\boldsymbol{H}_{22,i}^{-1}\boldsymbol{H}_{12,i}^{T}\right|
$$

where

$$
\boldsymbol{H}_{12,i} \equiv E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}
$$

$$
- \sum_{j=1}^{n_i}E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^{-1}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}^T
$$

and

$$
\boldsymbol{H}_{22,i} \equiv \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}
$$

$$
- \sum_{j=1}^{n_i}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^{-1}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}^T.
$$

A.16   Derivation of Algorithm 12

Algorithm 12 is the three-level counterpart of Algorithm 7 and they each use the same logic. Therefore, the Algorithm 12 follows from arguments similar to those given in Section A.7.

A.17   Derivation of Result 7

Note that

$$
\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}) \propto m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\, m_{(\boldsymbol{\beta},\boldsymbol{u})\to\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u})
$$

$$
= \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\text{L1}}\\ \text{vech}(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T)\\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T)\end{bmatrix}\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\text{L2}}\\ \text{vech}(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T)\\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L1}})^T)\\ \text{vec}(\boldsymbol{u}_{ij}^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T)\end{bmatrix}\end{bmatrix}\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}\right\}
$$

$$
= \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^T\boldsymbol{a} - \frac{1}{2}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^T\boldsymbol{A}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right\}
$$

where $\boldsymbol{a}$ and $\boldsymbol{A}$ are as given in Result 7. The last step uses facts such as $\text{vech}(\boldsymbol{M}) = \boldsymbol{D}_d^+\text{vec}(\boldsymbol{M})$ for any symmetric $d \times d$ matrix $\boldsymbol{M}$. Standard manipulations then lead to

$$\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}\boldsymbol{a} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}$$

and Result 7 then follows from extraction of the sub-blocks of $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{a}$ and the sub-blocks of $\boldsymbol{A}^{-1}$ corresponding to the non-zero positions of $\boldsymbol{A}$.

A.18   Derivation of Algorithm 13

As a function of $(\boldsymbol{\beta}, \boldsymbol{u})$, the logarithm of the fragment factor is:

$$\log \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{u}_{ij}^{\text{L2}}\|^2 + \text{const}$$

$$= (1/\sigma^2)\begin{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\\text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\end{bmatrix}\\\underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\text{L1}}\\\text{vech}(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T)\\\text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T)\end{bmatrix}\\\underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\underset{1\leq j\leq n_i}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\text{L2}}\\\text{vech}(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T)\\\text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L1}})^T)\\\text{vec}(\boldsymbol{u}_{ij}^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T)\end{bmatrix}\end{bmatrix}\end{bmatrix}^T\begin{bmatrix}\begin{bmatrix}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}^T\boldsymbol{y}_{ij}\\-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij})\end{bmatrix}\\\underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\sum_{j=1}^{n_i}(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{y}_{ij}\\-\frac{1}{2}\sum_{j=1}^{n_i}\boldsymbol{D}_{q_1}^T\text{vec}((\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{Z}_{ij}^{\text{L1}})\\-\sum_{j=1}^{n_i}\text{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\text{L1}})\end{bmatrix}\\\underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\underset{1\leq j\leq n_i}{\text{stack}}\begin{bmatrix}(\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{y}_{ij}\\-\frac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}((\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{Z}_{ij}^{\text{L2}})\\-\text{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\text{L2}})\\-\text{vec}((\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{Z}_{ij}^{\text{L2}})\end{bmatrix}\end{bmatrix}\end{bmatrix} + \text{const.}$$

Therefore, from equations (8) and (9) of Wand (2017),

$$m_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \longleftarrow \exp\left\{\begin{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\\text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\end{bmatrix}\\\underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\text{L1}}\\\text{vech}(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T)\\\text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T)\end{bmatrix}\\\underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\underset{1\leq j\leq n_i}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\text{L2}}\\\text{vech}(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T)\\\text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L1}})^T)\\\text{vec}(\boldsymbol{u}_{ij}^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T)\end{bmatrix}\end{bmatrix}\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}\right\}$$

where

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\,\boldsymbol{u})} \equiv \mu_{\mathsf{q}(1/\sigma^2)}
\begin{bmatrix}
\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}^{T}\boldsymbol{y}_{ij} \\[2em]
-\tfrac{1}{2}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{D}_{p}^{T}\mathrm{vec}(\boldsymbol{X}_{ij}^{T}\boldsymbol{X}_{ij}) \\[2em]
\underset{1\le i\le m}{\mathrm{stack}}
\begin{bmatrix}
\displaystyle\sum_{j=1}^{n_i}(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{y}_{ij} \\[1.5em]
-\tfrac{1}{2}\displaystyle\sum_{j=1}^{n_i}\boldsymbol{D}_{q_1}^{T}\mathrm{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L1}}) \\[1.5em]
-\displaystyle\sum_{j=1}^{n_i}\mathrm{vec}(\boldsymbol{X}_{ij}^{T}\boldsymbol{Z}_{ij}^{\mathrm{L1}})
\end{bmatrix} \\[6em]
\underset{1\le i\le m}{\mathrm{stack}}
\begin{bmatrix}
\underset{1\le j\le n_i}{\mathrm{stack}}
\begin{bmatrix}
(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^{T}\boldsymbol{y}_{ij} \\
-\tfrac{1}{2}\boldsymbol{D}_{q_2}^{T}\mathrm{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L2}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L2}}) \\
-\mathrm{vec}(\boldsymbol{X}_{ij}^{T}\boldsymbol{Z}_{ij}^{\mathrm{L2}}) \\
-\mathrm{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L2}})
\end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

and $\mu_{\mathsf{q}(1/\sigma^2)}$ denotes expectation of $1/\sigma^2$ with respect to the normalization of

$$
m_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\rightarrow\,\sigma^2}(\sigma^2)\,m_{\sigma^2\,\rightarrow\,\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)}(\sigma^2).
$$

This is an Inverse χ^2 density function with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\leftrightarrow\,\sigma^2}$ and, from Table S.1 in the online supplement of Wand (2017), we have

$$
\mu_{\mathsf{q}(1/\sigma^2)} \longleftarrow \left((\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\leftrightarrow\,\sigma^2})_1 + 1\right)/(\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\leftrightarrow\,\sigma^2})_2.
$$

The other factor to stochastic node message update is

$$
m_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\rightarrow\,\sigma^2}(\sigma^2) \longleftarrow \exp\left\{
\begin{bmatrix}
\log(\sigma^2) \\[0.5em]
1/\sigma^2
\end{bmatrix}^{T}
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\rightarrow\,\sigma^2}
\right\}
$$

where

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\rightarrow\,\sigma^2} \equiv
\begin{bmatrix}
-\tfrac{1}{2}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}o_{ij} \\[2em]
-\tfrac{1}{2}\displaystyle\sum_{i=1}^{m}E_{\mathsf{q}}\{\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{u}_{ij}^{\mathrm{L2}}\|^2\}
\end{bmatrix}
$$

with E_{q} denoting expectation with respect to the normalization of

$$
m_{\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\,\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\,m_{(\boldsymbol{\beta},\,\boldsymbol{u})\,\rightarrow\,\mathfrak{p}(\boldsymbol{y}\,|\,\boldsymbol{\beta},\,\boldsymbol{u},\,\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u}).
$$

Observing that

$$
E_{\mathsf{q}}\{\|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} - \boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{u}_{ij}^{\mathrm{L2}}\|^2\}
$$

$$
\begin{aligned}
&= \|\boldsymbol{y}_{ij} - \boldsymbol{X}_{ij}\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_{ij}^{\mathrm{L1}}\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} - \boldsymbol{Z}_{ij}^{\mathrm{L2}}\,\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\|^2 + \mathrm{tr}(\boldsymbol{X}_{ij}^{T}\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta})}) \\
&\quad + \mathrm{tr}\{(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\} + \mathrm{tr}\{(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\} \\
&\quad + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^{T}\boldsymbol{X}_{ij}E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^{T}\}\big] \\
&\quad + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^{T}\boldsymbol{X}_{ij}E_{\mathsf{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^{T}\}\big] \\
&\quad + 2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^{T}\boldsymbol{Z}_{ij}^{\mathrm{L1}}E_{\mathsf{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}} - \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^{T}\}\big]
\end{aligned}
$$

Result 7 shows how the sub-blocks of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\mid\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ are related to the required sub-vectors of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$. These matrices are obtained from $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}\mid\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ in the call to THREELEVELNATURALTOCOMMONPARAMETERS algorithm (Algorithm 12).

A.19 Derivation of Result 8

The derivation of Result 8 is very similar to that for Result 7.

A.20 Derivation of Algorithm 14

The logarithm on the fragment factor is, as a function of $(\boldsymbol{\beta},\boldsymbol{u})$:

$$
\log \mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\mid\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}) = \tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}) - \tfrac{1}{2}\sum_{i=1}^{m}(\boldsymbol{u}_i^{\text{L1}})^T(\boldsymbol{\Sigma}^{\text{L1}})^{-1}\boldsymbol{u}_i^{\text{L1}}
$$

$$
-\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\boldsymbol{u}_{ij}^{\text{L2}})^T(\boldsymbol{\Sigma}^{\text{L2}})^{-1}\boldsymbol{u}_{ij}^{\text{L2}} + \text{const}
$$

$$
= \begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i^{\text{L1}} \\ \text{vech}(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T) \\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T) \end{bmatrix} \\ \underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\underset{1\leq j\leq n_i}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_{ij}^{\text{L2}} \\ \text{vech}(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T) \\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L1}})^T) \\ \text{vec}(\boldsymbol{u}_{ij}^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T) \end{bmatrix}\end{bmatrix} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\ \underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix} \boldsymbol{0}_{q_1} \\ -\tfrac{1}{2}\boldsymbol{D}_{q_1}^T\text{vec}((\boldsymbol{\Sigma}^{\text{L1}})^{-1}) \\ \boldsymbol{0}_{pq_1} \end{bmatrix} \\ \underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\underset{1\leq j\leq n_i}{\text{stack}}\begin{bmatrix} \boldsymbol{0}_{q_2} \\ -\tfrac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}((\boldsymbol{\Sigma}^{\text{L2}})^{-1}) \\ \boldsymbol{0}_{pq_2} \\ \boldsymbol{0}_{q_1q_2} \end{bmatrix}\end{bmatrix} \end{bmatrix}.
$$

Therefore, from equations (8) and (9) of Wand (2017),

$$
m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\mid\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) \longleftarrow \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i^{\text{L1}} \\ \text{vech}(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T) \\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T) \end{bmatrix} \\ \underset{1\leq i\leq m}{\text{stack}}\begin{bmatrix}\underset{1\leq j\leq n_i}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_{ij}^{\text{L2}} \\ \text{vech}(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T) \\ \text{vec}(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L1}})^T) \\ \text{vec}(\boldsymbol{u}_{ij}^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T) \end{bmatrix}\end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}\mid\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\to(\boldsymbol{\beta},\boldsymbol{u})} \right\}
$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to(\boldsymbol{\beta},\boldsymbol{u})} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\[4pt] -\frac{1}{2}\boldsymbol{D}_p^T\mathrm{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\[6pt] \underset{1\le i\le m}{\mathrm{stack}}\begin{bmatrix} \boldsymbol{0}_{q_1} \\ -\frac{1}{2}\boldsymbol{D}_{q_1}^T\mathrm{vec}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})}\big) \\ \boldsymbol{0}_{pq_1} \end{bmatrix} \\[10pt] \underset{1\le i\le m}{\mathrm{stack}}\Bigg[\underset{1\le j\le n_i}{\mathrm{stack}}\begin{bmatrix} \boldsymbol{0}_{q_2} \\ -\frac{1}{2}\boldsymbol{D}_{q_2}^T\mathrm{vec}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})}\big) \\ \boldsymbol{0}_{pq_2} \\ \boldsymbol{0}_{q_1q_2} \end{bmatrix}\Bigg] \end{bmatrix}.$$

Here $\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})}$ denotes expectation of $(\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1}$ with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L1}}}(\boldsymbol{\Sigma}^{\mathrm{L1}})\,m_{\boldsymbol{\Sigma}^{\mathrm{L1}}\to\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})}(\boldsymbol{\Sigma}^{\mathrm{L1}}).$$

which is an Inverse G-Wishart density function with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\leftrightarrow\boldsymbol{\Sigma}^{\mathrm{L1}}}$ and, according to Table S.1 in the online supplement of Wand (2017), leads to

$$\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})} \longleftarrow \{\omega_{47}+\tfrac{1}{2}(q_1+1)\}\{\mathrm{vec}^{-1}(\boldsymbol{\omega}_{48})\}^{-1}$$

where ω_{47} is the first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\leftrightarrow\boldsymbol{\Sigma}^{\mathrm{L1}}}$ and $\boldsymbol{\omega}_{48}$ is the vector containing the remaining entries of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\leftrightarrow\boldsymbol{\Sigma}^{\mathrm{L1}}}$. The treatment of $\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})}$ is analogous.

The message from $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})$ to $\boldsymbol{\Sigma}^{\mathrm{L1}}$ is

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L1}}}(\boldsymbol{\Sigma}^{\mathrm{L1}})\longleftarrow\exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}^{\mathrm{L1}}| \\ \mathrm{vech}\big((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1}\big) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L1}}}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L1}}} \equiv \begin{bmatrix} -\frac{1}{2}\,m \\ -\frac{1}{2}\sum_{i=1}^{m}\boldsymbol{D}_{q_1}^T\mathrm{vec}[E_{\mathfrak{q}}\{\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_i^{\mathrm{L1}})^T\}] \end{bmatrix}$$

with $E_{\mathfrak{q}}$ denoting expectation with respect to the normalization of

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\,m_{(\boldsymbol{\beta},\boldsymbol{u})\to\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})}(\boldsymbol{\beta},\boldsymbol{u}).$$

Similarly, the message from $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})$ to $\boldsymbol{\Sigma}^{\mathrm{L2}}$ is

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L2}}}(\boldsymbol{\Sigma}^{\mathrm{L2}})\longleftarrow\exp\left\{\begin{bmatrix} \log|\boldsymbol{\Sigma}^{\mathrm{L2}}| \\ \mathrm{vech}\big((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1}\big) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L2}}}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\to\boldsymbol{\Sigma}^{\mathrm{L2}}} \equiv \begin{bmatrix} -\frac{1}{2}\sum_{i=1}^{m} n_i \\ -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{D}_{q_2}^T\mathrm{vec}[E_{\mathfrak{q}}\{\boldsymbol{u}_{ij}^{\mathrm{L2}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\}] \end{bmatrix}.$$

Now note that

$$E_{\mathsf{q}}\{\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_i^{\mathrm{L1}})^T\} = \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}^T + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \quad \text{and} \quad E_{\mathsf{q}}\{\boldsymbol{u}_{ij}^{\mathrm{L2}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\} = \boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}^T + \boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$$

where, similar to before, $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \equiv E_{\mathsf{q}}(\boldsymbol{u}_i^{\mathrm{L1}})$, $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \equiv \mathrm{Cov}_{\mathsf{q}}(\boldsymbol{u}_i^{\mathrm{L1}})$ and $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$ is defined similarly. Result 8 links sub-blocks of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\,\boldsymbol{u}\,|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\,\boldsymbol{\Sigma}^{\mathrm{L2}}) \leftrightarrow (\boldsymbol{\beta},\,\boldsymbol{u})}$ with the required sub-vectors of $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and sub-blocks of $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{\beta},\boldsymbol{u})}$. We then call upon Algorithm 12 to obtain $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_i^{\mathrm{L1}})}$, $1 \leq i \leq m$, as well as $\boldsymbol{\mu}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$ and $\boldsymbol{\Sigma}_{\mathsf{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$, $1 \leq i \leq m$, $1 \leq j \leq n_i$.

Algorithm 14 is a proceduralization of each of these results.