Journal of Machine Learning Research 21 (2020) 1-62 Submitted 3/19; Revised 7/20; Published 9/20

Streamlined Computing for Variational Inference with Higher Level Random Effects

Tui H. Nolan TUI.NOLAN@UTS.EDU.AU
School of Mathematical and Physical Sciences
University of Technology Sydney
P.O. Box 123, Broadway, 2007, Australia

Marianne Menictas MMENICTAS@GMAIL.COM
School of Mathematical and Physical Sciences
University of Technology Sydney
P.O. Box 123, Broadway, 2007, Australia

Matt P. Wand MATT.WAND@UTS.EDU.AU
School of Mathematical and Physical Sciences
University of Technology Sydney
P.O. Box 123, Broadway, 2007, Australia

Editor: Xiaotong Shen

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 and 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 and Bates, 2000), multilevel models (e.g. Goldstein, 2010), panel data analysis (e.g. Baltagi, 2013) and small area estimation (e.g. Rao and 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.

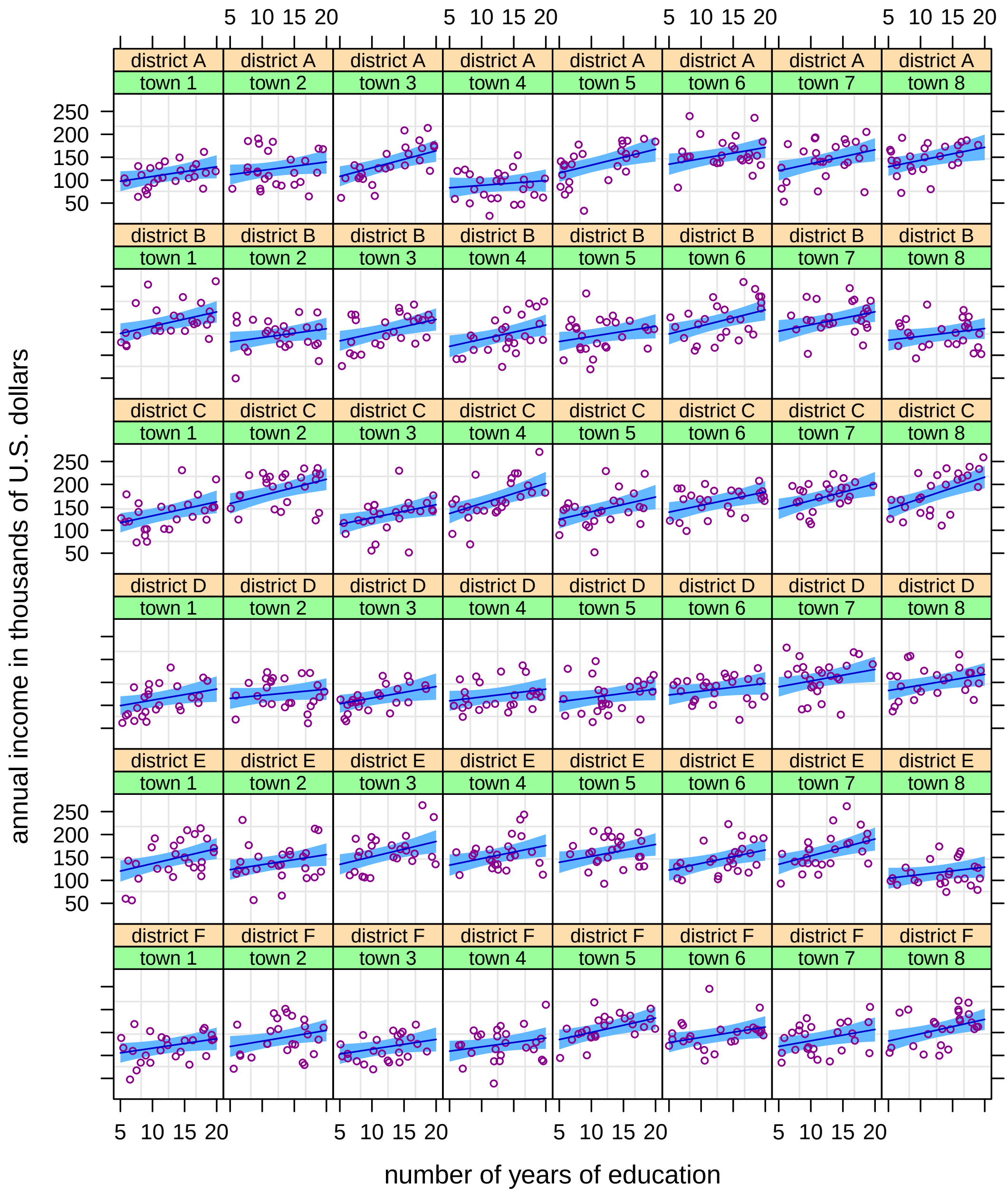

Figure 1: *Simulated three-level data according to 6 districts, each having 8 towns, each having 25 randomly chosen residents. In each panel, the line corresponds to a mean field variational Bayes fit, according to an appropriate multilevel model and the shaded region corresponds to pointwise 95% credible intervals for the mean response.*

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 and 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 mean field variational Bayes fit of a three-level random intercepts and slopes linear mixed model, as explained in Section 5.1. 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 and Wand (2020) 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 A.1–A.4 in Appendix A. 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 introducing the four fundamental algorithms in Section 3 and laying them out in Appendix A we then derive an additional eight algorithms, labeled Algorithms 1–8, that facilitate variational inference for two-level and three-level linear mixed models. The mean field variational Bayes approach is dealt with in Algorithms 1 and 5. The remaining six algorithms are concerned with streamlined factor graph fragment updates according to the variational message passing infrastructure described in Wand (2017). As explained in Section 3.2 there, the message passing approach has the advantage compartmentalization of variational inference algebra and code. Once a key fragment is identified, it only has to be derived and coded once and then can be used in models of arbitrarily large size. 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 1–8, but are be treated elsewhere (e.g. Menictas et al., 2020). 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 and Wand (2016). However, Lee and 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 summarize issues involving matrix algebra and point to Appendix A. This appendix presents four algorithms for solving higher level sparse matrix problems which are fundamental for variational inference involving general models with hierarchical random effects structure. Streamlined variational inference for mixed models possessing two-level random effects structure is treated in Section 4, followed by treatment of the three-level situation in Section 5. Derivations of all results and algorithms given in Sections 4 and 5 are deferred to Appendix B. Section 6 demonstrates the speed advantages of streamlining for variational inference in random effects models via some computational complexity calculations and 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 and Wand (2010) and Blei et al. (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)

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

for hyperparameters $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ 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 $\boldsymbol{A}_{\boldsymbol{G}}$ and $\boldsymbol{A}_{\boldsymbol{R}}$ with conjugate Inverse G-Wishart distributions (Wand, 2017). The prior specification of $\boldsymbol{G}$ and $\boldsymbol{R}$ involves the specifications

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

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).

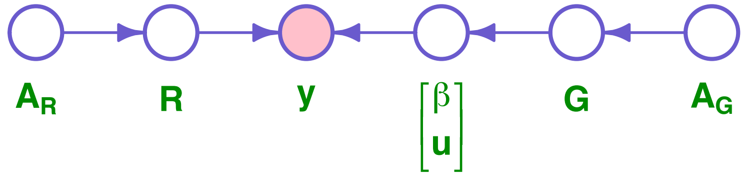

Figure 2: *Directed acyclic graph representation of model (1). The shading of the $\boldsymbol{y}$ node indicates that this vector of response values is observed.*

Full Bayesian inference for the $\boldsymbol{\beta}$, $\boldsymbol{G}$ and $\boldsymbol{R}$ and the random effects $\boldsymbol{u}$ involves the posterior density function $\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_{\boldsymbol{G}}, \boldsymbol{A}_{\boldsymbol{R}}, \boldsymbol{G}, \boldsymbol{R}|\boldsymbol{y})$, 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

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

for density functions $\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{A}_{\boldsymbol{G}}, \boldsymbol{A}_{\boldsymbol{R}})$ 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

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

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. Ormerod and Wand, 2010, Algorithm 1) 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

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

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 $\boldsymbol{A}_{\boldsymbol{G}}$ and $\boldsymbol{A}_{\boldsymbol{R}}$.

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.

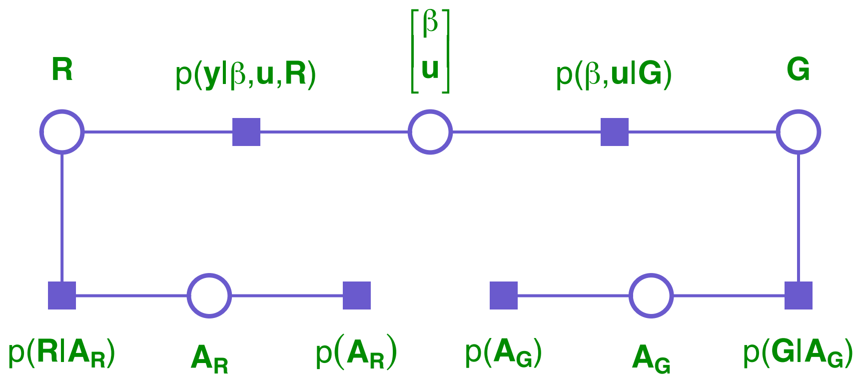

Figure 3: *Factor graph representation of the product structure of (6) with the solid rectangles corresponding to the factors and open circles corresponding to the unobserved random vectors and random matrices of the Bayesian linear mixed model given by (1) and (2), known as stochastic nodes. Edges join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

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

where

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

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 $d\times1$ Multivariate Normal random vector

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

this involves re-expression of its density function according to

$$\begin{aligned}\mathfrak{p}(\boldsymbol{x}) &= (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}\\ &= \exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta}-A(\boldsymbol{\eta})-\tfrac{d}{2}\log(2\pi)\}\end{aligned}$$

where

$$\boldsymbol{T}(\boldsymbol{x}) \equiv \begin{bmatrix}\boldsymbol{x}\\ \text{vech}(\boldsymbol{x}\,\boldsymbol{x}^T)\end{bmatrix} \quad\text{and}\quad \boldsymbol{\eta}\equiv\begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix}\equiv\begin{bmatrix}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ -\tfrac{1}{2}\,\boldsymbol{D}_d^T\text{vec}(\boldsymbol{\Sigma}^{-1})\end{bmatrix}$$

are, respectively, the *sufficient statistic* and *natural parameter* vectors. The matrix $\boldsymbol{D}_d$, 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 $\boldsymbol{D}_d\,\text{vech}(\boldsymbol{A})=\text{vec}(\boldsymbol{A})$ for any symmetric $d\times d$ matrix $\boldsymbol{A}$. The function

$$A(\boldsymbol{\eta}) = -\tfrac{1}{4}\,\boldsymbol{\eta}_1^T\left\{\text{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right\}^{-1}\boldsymbol{\eta}_1-\tfrac{1}{2}\log\left|-2\,\text{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right|$$

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

$$\boldsymbol{\mu}=-\tfrac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right\}^{-1}\boldsymbol{\eta}_1\quad\text{and}\quad\boldsymbol{\Sigma}=-\tfrac{1}{2}\left\{\text{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right\}^{-1}.$$

The vec and vech matrix operators are reasonably well-established (e.g. Gentle, 2007). If $\boldsymbol{a}$ is a $d^2\times 1$ vector then $\text{vec}^{-1}(\boldsymbol{a})$ is the $d\times d$ matrix such that $\text{vec}\big(\text{vec}^{-1}(\boldsymbol{a})\big)=\boldsymbol{a}$. We also require vec inversion of non-square matrices. If $\boldsymbol{a}$ is a $(d_1d_2)\times 1$ vector then $\text{vec}^{-1}_{d_1\times d_2}(\boldsymbol{a})$ is the $d_1\times d_2$ matrix such that $\text{vec}\big(\text{vec}^{-1}_{d_1\times d_2}(\boldsymbol{a})\big)=\boldsymbol{a}$.

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 and Massam, 2005; Maestrini and Wand, 2020). For any positive integer d, let G be an undirected graph with d nodes labeled $1,\ldots,d$ and set E consisting of sets of pairs of nodes that are connected by an edge. We say that the symmetric $d\times d$ matrix $\boldsymbol{M}$ *respects* G if

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

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

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

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

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

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

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

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

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

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

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

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 distribution. We write

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

for this Inverse-G-Wishart(G, ξ, λ) 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 and 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. Matrix Algebraic Background

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

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

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.

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 and Wand (2020), are similar to those used traditionally for fitting frequentist random effects (Pinheiro and 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.

Appendix A provides the details of the multilevel sparse matrix problem algorithms used in the upcoming variational inference algorithms. There are four such matrix algebraic algorithms:

SolveTwoLevelSparseMatrix	Algorithm A.1
SolveTwoLevelSparseLeastSquares	Algorithm A.2
SolveThreeLevelSparseMatrix	Algorithm A.3
SolveThreeLevelSparseLeastSquares	Algorithm A.4

We use these four descriptive names in the variational inference algorithms that begin in the next section.

4. Two-Level Models

We now present streamlined algorithms for two-level linear mixed models.

4.1 Mean Field Variational Bayes

Consider the following Bayesian model:

$$\begin{gathered}\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{u}_i,\sigma^2 \stackrel{\text{ind.}}{\sim} N(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}_i\,\boldsymbol{u}_i,\sigma^2\,\boldsymbol{I}),\quad \boldsymbol{u}_i|\boldsymbol{\Sigma} \stackrel{\text{ind.}}{\sim} N(\boldsymbol{0},\boldsymbol{\Sigma}),\quad 1\le i\le m,\\ \boldsymbol{\beta}\sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\quad \sigma^2|a_{\sigma^2}\sim \text{Inverse-}\chi^2(\nu_{\sigma^2},1/a_{\sigma^2}),\\ a_{\sigma^2}\sim\text{Inverse-}\chi^2(1,1/(\nu_{\sigma^2}s_{\sigma^2}^2)),\\ \boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}}\sim\text{Inverse-G-Wishart}\big(G_{\text{full}},\nu_{\boldsymbol{\Sigma}}+2q-2,\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big),\\ \boldsymbol{A}_{\boldsymbol{\Sigma}}\sim\text{Inverse-G-Wishart}(G_{\text{diag}},1,\boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}),\quad \boldsymbol{\Lambda}_{\boldsymbol{A}_{\boldsymbol{\Sigma}}}\equiv\{\nu_{\boldsymbol{\Sigma}}\text{diag}(s_{\boldsymbol{\Sigma},1}^2,\ldots,s_{\boldsymbol{\Sigma},q}^2)\}^{-1},\end{gathered}\tag{7}$$

where matrix dimensions, for $1\le i\le m$, are as follows:

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

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

$$\boldsymbol{y}\equiv\begin{bmatrix}\boldsymbol{y}_1\\ \vdots\\ \boldsymbol{y}_m\end{bmatrix},\ \boldsymbol{X}\equiv\begin{bmatrix}\boldsymbol{X}_1\\ \vdots\\ \boldsymbol{X}_m\end{bmatrix},\ \boldsymbol{Z}\equiv\underset{1\le i\le m}{\text{blockdiag}}(\boldsymbol{Z}_i),\ \boldsymbol{u}\equiv\begin{bmatrix}\boldsymbol{u}_1\\ \vdots\\ \boldsymbol{u}_m\end{bmatrix}\text{ and }\boldsymbol{C}\equiv[\boldsymbol{X}\ \boldsymbol{Z}].$$

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 (7) 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 and Wand (2013). As explained in Huang and Wand (2013), such priors allow standard deviation and correlation parameters to have arbitrary non-informativeness.

Now consider the following mean field restriction on the joint posterior density function of all parameters in (7):

$$\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u},a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}},\sigma^2,\boldsymbol{\Sigma}|\boldsymbol{y})\approx\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}})\,\mathfrak{q}(\sigma^2,\boldsymbol{\Sigma})\tag{8}$$

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.

Bishop, 2006, equation (10.9)) leads to the optimal $\mathfrak{q}$-density functions for the parameters of interest being as follows:

$$\begin{aligned} &\mathfrak{q}^*(\boldsymbol{\beta},\boldsymbol{u}) \text{ has a } N\big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}\big) \text{ density function,}\\ &\mathfrak{q}^*(\sigma^2) \text{ has an Inverse-}\chi^2\big(\xi_{\mathfrak{q}(\sigma^2)},\lambda_{\mathfrak{q}(\sigma^2)}\big) \text{ density function}\\ \text{and }\ &\mathfrak{q}^*(\boldsymbol{\Sigma}) \text{ has an Inverse-G-Wishart}(G_{\text{full}},\xi_{\mathfrak{q}(\boldsymbol{\Sigma})},\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}) \text{ density function.} \end{aligned} \tag{9}$$

The optimal $\mathfrak{q}$-density parameters are determined via an iterative coordinate ascent algorithm, with details deferred to Appendix B.2. Algorithm 2 of Lee and Wand (2016) is a naïve mean field variational Bayes algorithm for a class of two-level Gaussian response linear mixed models that includes model (7) as a special case. Subsequent algorithms in Lee and Wand (2016) achieve streamlining. In the current article, we offer an alternative approach, based on Algorithms 1 and 5, 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

$$\begin{aligned} &\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C}+\boldsymbol{D}_{\text{MFVB}})^{-1}(\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{y}+\boldsymbol{o}_{\text{MFVB}})\\ \text{and}\quad &\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C}+\boldsymbol{D}_{\text{MFVB}})^{-1} \end{aligned} \tag{10}$$

where

$$\boldsymbol{R}_{\text{MFVB}} \equiv \mu_{\mathfrak{q}(1/\sigma^2)}^{-1}\boldsymbol{I},\ \boldsymbol{D}_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{I}_m\otimes\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\end{bmatrix} \text{ and } \boldsymbol{o}_{\text{MFVB}} \equiv \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\\ \boldsymbol{0}\end{bmatrix}. \tag{11}$$

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 $\boldsymbol{\Sigma}$:

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} &= \text{top left-hand } p\times p \text{ sub-block of } (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C}+\boldsymbol{D}_{\text{MFVB}})^{-1},\\ \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} &= \text{subsequent } q\times q \text{ diagonal sub-blocks of } (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C}+\boldsymbol{D}_{\text{MFVB}})^{-1}\\ &\quad\ \text{below } \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})},\ 1\le i\le m, \text{ and}\\ E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\beta)})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\} &= \text{subsequent } p\times q \text{ sub-blocks of}\\ &\quad (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C}+\boldsymbol{D}_{\text{MFVB}})^{-1} \text{ to the right of } \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})},\ 1\le i\le m. \end{aligned} \tag{12}$$

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

Result 1 *The mean field variational Bayes updates of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and each of the sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ listed in (12) are expressible as a two-level sparse matrix least squares problem (see Appendix A.1) 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 (29), 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, according to the notation in (27) and (28),*

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

and

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

Result 1 implies that the SolveTwoLevelSparseLeastSquares algorithm listed in Algorithm A.2 applies for handling the $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ sub-block updates. A derivation is in Appendix B.1. This results in Algorithm 1 for streamlined mean field variational Bayes for the two-level Gaussian response linear mixed model. A derivation is given in Appendix B.2.

An important aspect of Result 1 and Algorithm 1 is that the vector $(\boldsymbol{\beta}, \boldsymbol{u})$ is treated as an entity in the updates. This contrasts with block Markov chain Monte Carlo sampling schemes where sub-vectors of $(\boldsymbol{\beta}, \boldsymbol{u})$ are updated separately. In the case of variational inference, block updating of the sub-vectors of $(\boldsymbol{\beta}, \boldsymbol{u})$ corresponds to the imposition of more stringent product restrictions on the $\mathfrak{q}$-density of $(\boldsymbol{\beta}, \boldsymbol{u})$ and degradation of accuracy.

Algorithm 1 uses the mean field variational Bayes approximate marginal log-likelihood $\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\}$ in its stopping criterion. For model (7) this is given by

$$\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\} = E_{\mathfrak{q}}\{\log\, \mathfrak{p}(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a_{\sigma^2}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}}) - \mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a_{\sigma^2}, \boldsymbol{\Sigma}, \boldsymbol{A}_{\boldsymbol{\Sigma}})\}. \tag{13}$$

An explicit streamlined expression for $\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\}$ and corresponding derivation is given in Nolan and Wand (2020).

4.2 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 (7) admits the following factorization:

$$\mathfrak{p}(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, \boldsymbol{\Sigma}, a_{\sigma^2}, \boldsymbol{A}_{\boldsymbol{\Sigma}}) = \mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2)\mathfrak{p}(\sigma^2|a_{\sigma^2})\mathfrak{p}(a_{\sigma^2})\mathfrak{p}(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{\Sigma})\mathfrak{p}(\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}})\mathfrak{p}(\boldsymbol{A}_{\boldsymbol{\Sigma}}). \tag{14}$$

Figure 4 shows a factor graph representation of (14) with color-coding of *fragment* types, according to the nomenclature in Wand (2017).

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 brown in Figure 4, are given in simple naïve forms in Wand (2017) without matrix algebraic streamlining. The next two subsections overcome this deficiency.

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

$$p + mq + \tfrac{1}{2}(p + mq)(p + mq + 1) \tag{15}$$

Algorithm 1 *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 (7) with product density restriction (8).*

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

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

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

Cycle:

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

$$\boldsymbol{b}_i \longleftarrow \begin{bmatrix} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2} \boldsymbol{y}_i \\ m^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix}, \ \boldsymbol{B}_i \longleftarrow \begin{bmatrix} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2} \boldsymbol{X}_i \\ m^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2} \\ \boldsymbol{0} \end{bmatrix}, \ \overset{\bullet}{\boldsymbol{B}}_i \longleftarrow \begin{bmatrix} \mu_{\mathfrak{q}(1/\sigma^2)}^{1/2} \boldsymbol{Z}_i \\ \boldsymbol{0} \\ \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}^{1/2} \end{bmatrix}.$$

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

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

$\lambda_{\mathfrak{q}(\sigma^2)} \longleftarrow \mu_{\mathfrak{q}(1/a_{\sigma^2})} \;\; ; \;\; \Lambda_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1})}$

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_1$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_1$

$E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_1$

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

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

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

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

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

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

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

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

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

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

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.

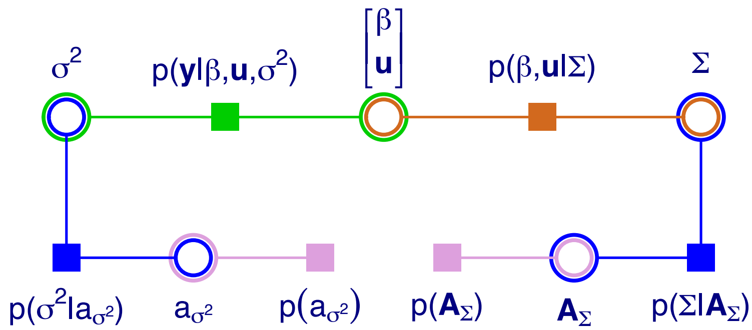

Figure 4: *Factor graph representation of the Bayesian two-level Gaussian response linear mixed model (7).*

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:

$$\begin{aligned}\exp\Big\{&\eta_{\beta_0}\beta_0+\eta_{\beta_1}\beta_1+\sum_{i=1}^m(\eta_{u_{i0}}u_{i0}+\eta_{u_{i1}}u_{i1})+\eta_{\beta_0^2}\beta_0^2+\eta_{\beta_1^2}\beta_1^2+\sum_{i=1}^m(\eta_{u_{i0}^2}\,u_{i0}^2+\eta_{u_{i1}^2}u_{i1}^2)\\&+\sum_{i=1}^m(\eta_{\beta_0u_{i0}}\beta_0\,u_{i0}+\eta_{\beta_0u_{i1}}\beta_0\,u_{i1}+\eta_{\beta_1u_{i0}}\beta_1\,u_{i0}+\eta_{\beta_1u_{i1}}\beta_1\,u_{i1})\Big\}.\end{aligned}\tag{16}$$

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 (15) when the number of groups is large. The reduced exponential family has 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 and 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, $\frac{1}{2}(p+mq)(p+mq-1)$ and $\frac{1}{2}\,p(p-1)+m\{\frac{1}{2}\,q(q-1)+pq\}$. 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.

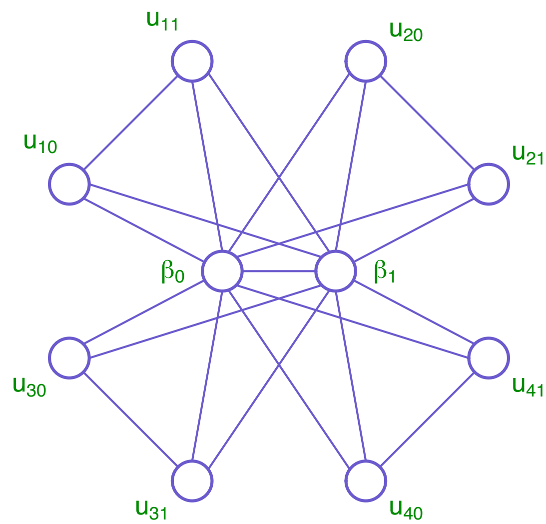

Figure 5: *Undirected probabilistic graph with edges coding the conditional dependencies of the entries of* $(\boldsymbol{\beta}, \boldsymbol{u})$ *given the rest for the case* $p = q = 2$ *and* $m = 4$.

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

$$m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta}, \boldsymbol{u}) = \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}\right\} \tag{17}$$

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

$$p + \tfrac{1}{2}p(p+1) + m\{q + \tfrac{1}{2}q(q+1) + pq\}. \tag{18}$$

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 (17) with natural parameter vector $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u})\,\to\,\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$ also of length (18).

Result 2 *The variational message passing updates of the quantities* $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}$, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}$, $1 \leq i \leq m$, *and the sub-blocks of* $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ *listed in (12) with* $\mathfrak{q}$*-density expectations with respect to the normalization of*

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

are expressible as a two-level sparse matrix problem (see Appendix A.1) with

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

and

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

is the partitioning of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ *that defines* $\boldsymbol{\eta}_{1,1}$, $\boldsymbol{\eta}_{1,2}$ *and* $\{(\boldsymbol{\eta}_{2,1,i},\boldsymbol{\eta}_{2,2,i},\boldsymbol{\eta}_{2,3,i}) : 1 \leq i \leq m\}$*. The solutions, according to the notation in (27) and (28), are* $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} = \boldsymbol{x}_1$, $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$ *and*

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

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 $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ to the corresponding $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ vector and important sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$, as illustrated by Result 2, 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 2.

It is easily shown (Appendix B.5) that messages between $\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)$ and σ^2 have Inverse Chi-Squared forms. For example,

$$m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix} 1/\sigma^2 \\ \log(\sigma^2) \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,\sigma^2}\right\}. \tag{19}$$

Algorithm 3 lists parameter updates for the two-level Gaussian likelihood fragment with streamlining according to the restricted exponential family form (17). Note that it makes use of SolveTwoLevelSparseMatrix (Algorithm A.1) since the natural parameter updates correspond to a two-level sparse matrix problem *without* least squares representation. Appendix B.5 provides details on the derivation of Algorithm 3.

As in Wand (2017), Algorithm 3 uses the notation

$$\boldsymbol{\eta}_{f\leftrightarrow\theta} \equiv \boldsymbol{\eta}_{f\rightarrow\theta} + \boldsymbol{\eta}_{\theta\rightarrow f}. \tag{20}$$

Algorithm 2 *The* TwoLevelNaturalToCommonParameters *algorithm for conversion of a two-level reduced natural parameter vector to its corresponding common parameters.*

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

$\boldsymbol{\omega}_1 \longleftarrow$ first p entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$

$\boldsymbol{\omega}_2 \longleftarrow$ next $\frac{1}{2}p(p+1)$ entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$; $\boldsymbol{\Omega}_3 \longleftarrow -2\text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\omega}_2)$

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

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

- $\boldsymbol{\omega}_{4i} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + \frac{1}{2}q(q+1) - 1$
- $\boldsymbol{\omega}_5 \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq - 1$
- $\boldsymbol{\omega}_6 \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q - 1$
- $\boldsymbol{\Omega}_{7i} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_q^{+T}\boldsymbol{\omega}_5)$; $\boldsymbol{\Omega}_{8i} \longleftarrow -\text{vec}^{-1}_{p\times q}(\boldsymbol{\omega}_6)$

$\mathcal{S}_2 \longleftarrow$ SolveTwoLevelSparseMatrix$\Big(\boldsymbol{\omega}_1, \boldsymbol{\Omega}_3, \big\{(\boldsymbol{\omega}_{4i}, \boldsymbol{\Omega}_{7i}, \boldsymbol{\Omega}_{8i}) : 1 \le i \le m\big\}\Big)$

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

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

- $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_2$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_2$
- $E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\}(\boldsymbol{u}_i - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_2$

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

4.4 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 brown in Figure 4.

As shown in Appendix B.7, the message from $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})$ to $(\boldsymbol{\beta},\boldsymbol{u})$ has the generic form (16) 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 (17). This implies that

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

with natural parameter vector $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ also of length (18). The reverse message has an analogous form.

Algorithm 3 *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: $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}$, $\boldsymbol{\eta}_{(\boldsymbol{\beta},\boldsymbol{u})\to\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$, $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to\sigma^2}$, $\boldsymbol{\eta}_{\sigma^2\to\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}$

Updates:

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

$$\mathcal{S}_3\longleftarrow\textsc{TwoLevelNaturalToCommonParameters}\Big(p,q,m,\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})}\Big)$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\text{ component of }\mathcal{S}_3\ \ ;\ \ \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\text{ component of }\mathcal{S}_3$$

$$\boldsymbol{\omega}_9\longleftarrow\boldsymbol{0}_p\ \ ;\ \ \boldsymbol{\omega}_{10}\longleftarrow\boldsymbol{0}_{\frac{1}{2}p(p+1)}\ \ ;\ \ \boldsymbol{\omega}_{11}\longleftarrow 0$$

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

$$\boldsymbol{\omega}_9\longleftarrow\boldsymbol{\omega}_9+\boldsymbol{X}_i^T\boldsymbol{y}_i\quad;\quad\boldsymbol{\omega}_{10}\longleftarrow\boldsymbol{\omega}_{10}-\tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{X}_i^T\boldsymbol{X}_i)$$

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\longleftarrow\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\text{ component of }\mathcal{S}_3\quad;\quad\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\longleftarrow\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\text{ component of }\mathcal{S}_3$$

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

$$\boldsymbol{\omega}_{11}\longleftarrow\boldsymbol{\omega}_{11}-\tfrac{1}{2}\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_i\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\|^2$$

$$\begin{aligned}\boldsymbol{\omega}_{11}\longleftarrow\ &\boldsymbol{\omega}_{11}-\tfrac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\boldsymbol{X}_i^T\boldsymbol{X}_i)-\tfrac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{Z}_i^T\boldsymbol{Z}_i)\\&-\text{tr}[\{\boldsymbol{Z}_i^T\boldsymbol{X}_iE_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)})^T\}]\end{aligned}$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to(\boldsymbol{\beta},\boldsymbol{u})}\longleftarrow\mu_{\mathfrak{q}(1/\sigma^2)}\begin{bmatrix}\boldsymbol{\omega}_9\\\boldsymbol{\omega}_{10}\\\underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{Z}_i^T\boldsymbol{y}_i\\-\frac{1}{2}\boldsymbol{D}_q^T\text{vec}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)\\-\text{vec}(\boldsymbol{X}_i^T\boldsymbol{Z}_i)\end{bmatrix}\end{bmatrix}$$

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\to\sigma^2}\longleftarrow\begin{bmatrix}-\frac{1}{2}\sum_{i=1}^m n_i\\\boldsymbol{\omega}_{11}\end{bmatrix}$$

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

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

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

are expressible as a two-level sparse matrix problem (see Appendix A.1) with

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

and

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

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* $\{(\boldsymbol{\eta}_{2,1,i},\boldsymbol{\eta}_{2,2,i},\boldsymbol{\eta}_{2,3,i}) : 1\le i\le m\}$. *The solutions are, according to the notation in (27) and (28),*

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

As shown in Appendix B.5, the message from $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})$ to $\boldsymbol{\Sigma}$ has the Inverse-G-Wishart form

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

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 4 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 (20). See Appendix B.5 for its derivation.

4.5 $\mathfrak{q}$-Density Determination After Variational Message Passing Convergence

After convergence of the variational message passing iterations, determination of $\mathfrak{q}$-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 (7) and its Figure 4 factor graph representation.

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

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

$$\mathcal{S}_A \longleftarrow \textsc{TwoLevelNaturalToCommonParameters}\Big(p,q,m,\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}\Big)$$

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

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

Algorithm 4 *The inputs, updates and outputs of the matrix algebraic streamlined Gaussian penalization fragment for two-level models.*

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

$\omega_{12}\longleftarrow$ first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\leftrightarrow\boldsymbol{\Sigma}}$; $\boldsymbol{\omega}_{13}\longleftarrow$ remaining entries of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\leftrightarrow\boldsymbol{\Sigma}}$

$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})}\longleftarrow\big\{\omega_{12}+\tfrac{1}{2}(q+1)\big\}\{\text{vec}^{-1}\big(\boldsymbol{D}_q^{+T}\boldsymbol{\omega}_{13}\big)\}^{-1}$

$\mathcal{S}_4\longleftarrow$ TwoLevelNaturalToCommonParameters$\Big(p,q,m,\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})}\Big)$

$\boldsymbol{\omega}_{14}\longleftarrow\mathbf{0}_{\frac{1}{2}q(q+1)}$

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

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

$\boldsymbol{\omega}_{14}\longleftarrow\boldsymbol{\omega}_{14}-\tfrac{1}{2}\boldsymbol{D}_q^T\text{vec}\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}\Big)$

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

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\to\boldsymbol{\Sigma}}\longleftarrow\begin{bmatrix}-\tfrac{1}{2}m\\ \boldsymbol{\omega}_{14}\end{bmatrix}$$

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

---

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

The error variance σ^2 has its optimal $\mathfrak{q}$-density function being that of an

$$\text{Inverse-}\chi^2\big(\xi_{\mathfrak{q}(\sigma^2)},\lambda_{\mathfrak{q}(\sigma^2)}\big)$$

distribution, and its parameters are determined from the steps:

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

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

where $\big(\boldsymbol{\eta}_{\mathfrak{q}(\sigma^2)}\big)_j$ 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$\big(G_{\text{full}},\xi_{\mathfrak{q}(\boldsymbol{\Sigma})},\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})}\big)$ distribution. The steps for determining its parameters after variational message passing convergence are:

$$
\begin{aligned}
&\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\,\rightarrow\,\boldsymbol{\Sigma}} + \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}})\rightarrow\boldsymbol{\Sigma}}\\
&\xi_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow -2\big(\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}\big)_1 - 2, \quad ; \quad \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma})} \longleftarrow -2\,\mathrm{vec}^{-1}\Big(\boldsymbol{D}_q^{+T}\big(\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\Sigma})}\big)_2\Big)
\end{aligned}
$$

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

4.6 Generalized Linear Mixed Model Extensions

In this article we focus on Gaussian response linear mixed models. The general principles also apply to non-Gaussian response models within the generalized linear mixed models framework. For the variational message passing approach Algorithm 4 is applicable for generalized linear mixed models as well since it involves nodes of the factor graph that are isolated from the likelihood factor. However, Algorithm 3 is specific to the Gaussian likelihood factor and extension to non-Gaussian likelihood cases is the subject of ongoing research.

5. Three-Level Models

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

5.1 Mean Field Variational Bayes

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

$$
\begin{gathered}
\boldsymbol{y}_{ij}|\boldsymbol{\beta},\boldsymbol{u}_i^{\mathrm{L1}},\boldsymbol{u}_{ij}^{\mathrm{L2}},\sigma^2 \overset{\text{ind.}}{\sim} N(\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}},\sigma^2\,\boldsymbol{I}),\\
\left[\begin{array}{c}\boldsymbol{u}_i^{\mathrm{L1}}\\ \boldsymbol{u}_{ij}^{\mathrm{L2}}\end{array}\right]\Big|\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}} \overset{\text{ind.}}{\sim} N\left(\left[\begin{array}{c}\boldsymbol{0}\\ \boldsymbol{0}\end{array}\right],\left[\begin{array}{cc}\boldsymbol{\Sigma}^{\mathrm{L1}} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{\Sigma}^{\mathrm{L2}}\end{array}\right]\right),\quad 1\le i\le m,\ 1\le j\le 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)),\\
\boldsymbol{\Sigma}^{\mathrm{L1}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}\sim\text{Inverse-G-Wishart}\big(G_{\text{full}},\nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}}+2q_1-2,(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})^{-1}\big),\\
\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}\sim\text{Inverse-G-Wishart}(G_{\text{diag}},1,\{\nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}}\text{diag}(s^2_{\boldsymbol{\Sigma}^{\mathrm{L1}},1},\ldots,s^2_{\boldsymbol{\Sigma}^{\mathrm{L1}},q_1})\}^{-1}),\\
\boldsymbol{\Sigma}^{\mathrm{L2}}|\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}\sim\text{Inverse-G-Wishart}\big(G_{\text{full}},\nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}}+2q_2-2,(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})^{-1}\big),\\
\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}\sim\text{Inverse-G-Wishart}(G_{\text{diag}},1,\{\nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}}\text{diag}(s^2_{\boldsymbol{\Sigma}^{\mathrm{L2}},1},\ldots,s^2_{\boldsymbol{\Sigma}^{\mathrm{L2}},q_2})\}^{-1}).
\end{gathered}
\tag{21}
$$

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

Algorithm 5 *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 (21) with product density restriction (22). The algorithm description requires more than one page and is continued on a subsequent page.*

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

Hyperparameter Inputs: $\boldsymbol{\mu}_{\boldsymbol{\beta}}(p\times 1)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}(p\times p)$ 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: $\mu_{\mathfrak{q}(1/\sigma^2)}>0$, $\mu_{\mathfrak{q}(1/a_{\sigma^2})}>0$, $\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})}(q_1\times q_1),\ \boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})}(q_2\times q_2),$

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

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

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

Cycle:

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

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

$$\boldsymbol{b}_{ij}\longleftarrow\begin{bmatrix}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\boldsymbol{y}_{ij}\\ \Big(\sum_{i=1}^m n_i\Big)^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\boldsymbol{\mu}_{\boldsymbol{\beta}}\\ \boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}\ \ ;\ \ \boldsymbol{B}_{ij}\longleftarrow\begin{bmatrix}\mu^{1/2}_{\mathfrak{q}(1/\sigma^2)}\boldsymbol{X}_{ij}\\ \Big(\sum_{i=1}^m n_i\Big)^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1/2}\\ \boldsymbol{O}\\ \boldsymbol{O}\end{bmatrix},$$

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

$\mathcal{S}_5\longleftarrow$ SolveThreeLevelSparseLeastSquares
$\Big(\big\{(\boldsymbol{b}_{ij},\boldsymbol{B}_{ij},\dot{\boldsymbol{B}}_{ij},\ddot{\boldsymbol{B}}_{ij}):1\le i\le m,1\le j\le n_i\big\}\Big)$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{x}_1$ component of $\mathcal{S}_5$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{A}^{11}$ component of $\mathcal{S}_5$

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

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}^{\mathrm{L1}}_i)}\longleftarrow\boldsymbol{x}_{2,i}$ component of $\mathcal{S}_5$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}^{\mathrm{L1}}_i)}\longleftarrow\boldsymbol{A}^{22,i}$ component of $\mathcal{S}_5$

$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}^{\mathrm{L1}}_i-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}^{\mathrm{L1}}_i)})^T\}\longleftarrow\boldsymbol{A}^{12,i}$ component of $\mathcal{S}_5$

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

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}^{\mathrm{L2}}_{ij})}\longleftarrow\boldsymbol{x}_{2,ij}$ component of $\mathcal{S}_5$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}^{\mathrm{L2}}_{ij})}\longleftarrow\boldsymbol{A}^{22,ij}$ component of $\mathcal{S}_5$

continued on a subsequent page ...

Algorithm 5 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\longleftarrow \boldsymbol{A}^{12,ij}\text{ component of }\mathcal{S}_5$$
$$E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\longleftarrow \boldsymbol{A}^{12,i,j}\text{ component of }\mathcal{S}_5$$
$$\lambda_{\mathfrak{q}(\sigma^2)}\longleftarrow\lambda_{\mathfrak{q}(\sigma^2)}+\left\|\boldsymbol{y}_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}-\boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\right\|^2$$
$$\lambda_{\mathfrak{q}(\sigma^2)}\longleftarrow\lambda_{\mathfrak{q}(\sigma^2)}+\text{tr}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})})+\text{tr}\{(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\}$$
$$\lambda_{\mathfrak{q}(\sigma^2)}\longleftarrow\lambda_{\mathfrak{q}(\sigma^2)}+\text{tr}((\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})$$
$$\lambda_{\mathfrak{q}(\sigma^2)}\longleftarrow\lambda_{\mathfrak{q}(\sigma^2)}+2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}\big]$$
$$\lambda_{\mathfrak{q}(\sigma^2)}\longleftarrow\lambda_{\mathfrak{q}(\sigma^2)}+2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\big]$$
$$\lambda_{\mathfrak{q}(\sigma^2)}\longleftarrow\lambda_{\mathfrak{q}(\sigma^2)}+2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{Z}_{ij}^{\text{L2}}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\big]$$
$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}\longleftarrow\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}+\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$$

$$\mu_{\mathfrak{q}(1/\sigma^2)}\longleftarrow\xi_{\mathfrak{q}(\sigma^2)}/\lambda_{\mathfrak{q}(\sigma^2)}$$
$$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L1}})}\longleftarrow(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L1}})}-q_1+1)\,\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L1}})}^{-1}\quad;\quad\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}\longleftarrow(\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}-q_2+1)\,\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\text{L2}})}^{-1}$$
$$\lambda_{\mathfrak{q}(a_{\sigma^2})}\longleftarrow\mu_{\mathfrak{q}(1/\sigma^2)}+1/(\nu_{\sigma^2}s_{\sigma^2}^2)\quad;\quad\mu_{\mathfrak{q}(1/a_{\sigma^2})}\longleftarrow\xi_{\mathfrak{q}(a_{\sigma^2})}/\lambda_{\mathfrak{q}(a_{\sigma^2})}$$
$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}})}\longleftarrow\text{diag}\big\{\text{diagonal}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})}\big)\big\}+\{\nu_{\boldsymbol{\Sigma}^{\text{L1}}}\,\text{diag}(s_{\boldsymbol{\Sigma}^{\text{L1}},1}^2,\ldots,s_{\boldsymbol{\Sigma}^{\text{L1}},q_1}^2)\}^{-1}$$
$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})}\longleftarrow\text{diag}\big\{\text{diagonal}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}\big)\big\}+\{\nu_{\boldsymbol{\Sigma}^{\text{L2}}}\,\text{diag}(s_{\boldsymbol{\Sigma}^{\text{L2}},1}^2,\ldots,s_{\boldsymbol{\Sigma}^{\text{L2}},q_2}^2)\}^{-1}$$
$$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}}^{-1})}\longleftarrow\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}})}\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}})}^{-1}\quad;\quad\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}}^{-1})}\longleftarrow\xi_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})}\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})}^{-1}.$$

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

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

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

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

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

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

$$\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u},a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}},\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}},\sigma^2,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}|\boldsymbol{y})\approx\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u},a_{\sigma^2},\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L1}}},\boldsymbol{A}_{\boldsymbol{\Sigma}^{\text{L2}}})\,\mathfrak{q}(\sigma^2,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}}).\tag{22}$$

The optimal $\mathfrak{q}$-densities have forms analogous to those given in (9) 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 $\mathfrak{q}^*(\boldsymbol{\Sigma}^{\text{L2}})$.

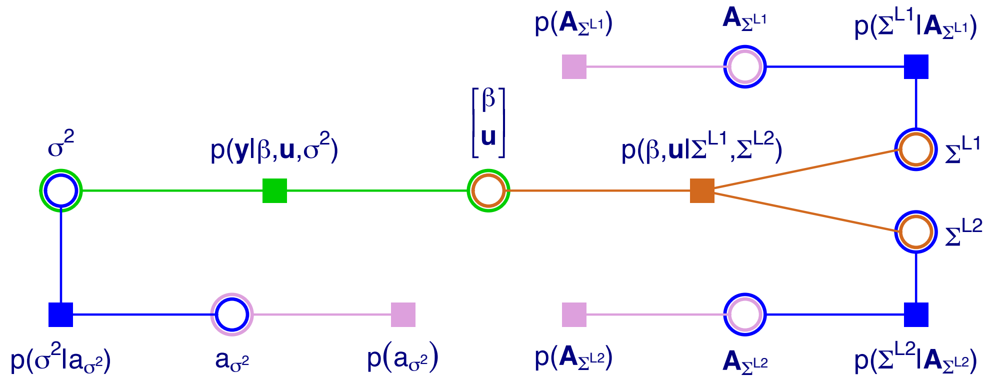

Figure 6: *Factor graph representation of the Bayesian three-level Gaussian response linear mixed model (21).*

As in the two-level case, only the following relatively small sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ are required for variational inference concerning σ^2, $\boldsymbol{\Sigma}^{\mathrm{L1}}$ and $\boldsymbol{\Sigma}^{\mathrm{L2}}$:

$$\begin{gathered}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})},\quad \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})},\quad \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})},\quad E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^T\}\\ E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\mathrm{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}\quad\text{and}\quad E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}\end{gathered}\tag{23}$$

for $1\le i\le m$ and $1\le j\le n_i$. Result 4 is the three-level analog of Result 1 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 4 *The mean field variational Bayes updates of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and each of the sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ corresponding to (23) are expressible as a three-level sparse matrix least squares problem (see Appendix A.2) 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 given by (35), are for $1\le j\le n_i,\ 1\le i\le m$:

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

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

with each of these matrices having $\tilde{o}_{ij} = o_{ij} + p + q_1 + q_2$ rows. The solutions are, according to notation illustrated by (30)–(32),

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

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

and

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

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

Algorithm 5 provides a streamlined mean field variational Bayes algorithm for approximate fitting and inference for (21). An explicit streamlined expression for the stopping criterion, $\log\{\underline{\mathfrak{p}}(\boldsymbol{y};\mathfrak{q})\}$, is given in Nolan and Wand (2020). We are not aware of any previously published variational inference algorithms that achieve streamlined inference for mixed models with three-level random effects.

5.2 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 (21) can be factorized as follows:

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

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

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 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 7 and 8. However, they each rely on the ThreeLevelNaturalToCommonParameters algorithm, which is listed as Algorithm 6.

Algorithm 6 *The* ThreeLevelNaturalToCommonParameters *algorithm. The algorithm description requires more than one page and is continued on a subsequent page.*

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

$\boldsymbol{\omega}_{15} \longleftarrow$ first p entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$

$\boldsymbol{\omega}_{16} \longleftarrow$ next $\frac{1}{2}\,p(p+1)$ entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$; $\boldsymbol{\Omega}_{17} \longleftarrow -2\text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\omega}_{16})$

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

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

- $\boldsymbol{\omega}_{18i} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + \frac{1}{2}\,q_1(q_1+1) - 1$
- $\boldsymbol{\omega}_{19} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq_1 - 1$
- $\boldsymbol{\omega}_{20} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
- $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_1 - 1$
- $\boldsymbol{\Omega}_{21i} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\omega}_{19})$; $\boldsymbol{\Omega}_{22i} \longleftarrow -\text{vec}^{-1}_{p\times q_1}(\boldsymbol{\omega}_{20})$

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

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

- For $j = 1,\ldots,n_i$:
  - $\boldsymbol{\omega}_{23ij} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
  - $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + \frac{1}{2}\,q_2(q_2+1) - 1$
  - $\boldsymbol{\omega}_{24} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
  - $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + pq_2 - 1$
  - $\boldsymbol{\omega}_{25} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
  - $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_1q_2 - 1$
  - $\boldsymbol{\omega}_{26} \longleftarrow$ sub-vector of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ with entries i_{stt} to i_{end} inclusive
  - $i_{\text{stt}} \longleftarrow i_{\text{end}} + 1$; $i_{\text{end}} \longleftarrow i_{\text{stt}} + q_2 - 1$
  - $\boldsymbol{\Omega}_{27ij} \longleftarrow -2\,\text{vec}^{-1}(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\omega}_{24})$; $\boldsymbol{\Omega}_{28ij} \longleftarrow -\text{vec}^{-1}_{p\times q_2}(\boldsymbol{\omega}_{25})$
  - $\boldsymbol{\Omega}_{29ij} \longleftarrow -\text{vec}^{-1}_{q_1\times q_2}(\boldsymbol{\omega}_{26})$

$$\mathcal{S}_6 \longleftarrow \textsc{SolveThreeLevelSparseMatrix}\Big(\boldsymbol{\omega}_{15}, \boldsymbol{\Omega}_{17}, \big\{(\boldsymbol{\omega}_{18i}, \boldsymbol{\Omega}_{21i}, \boldsymbol{\Omega}_{22i}) : 1 \le i \le m,\ (\boldsymbol{\omega}_{23ij}, \boldsymbol{\Omega}_{27ij}, \boldsymbol{\Omega}_{28ij}, \boldsymbol{\Omega}_{29ij}) : 1 \le i \le m, 1 \le j \le n_i\big\}\Big)$$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{x}_1$ component of $\mathcal{S}_6$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} \longleftarrow \boldsymbol{A}^{11}$ component of $\mathcal{S}_6$

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

- $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{x}_{2,i}$ component of $\mathcal{S}_6$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})} \longleftarrow \boldsymbol{A}^{22,i}$ component of $\mathcal{S}_6$,
- $E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})\}(\boldsymbol{u}_i^{\text{L1}} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\} \longleftarrow \boldsymbol{A}^{12,i}$ component of $\mathcal{S}_6$

continued on a subsequent page . . .

Algorithm 6 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

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

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})} \longleftarrow \boldsymbol{x}_{2,ij} \text{ component of } \mathcal{S}_6 \quad ; \quad \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})} \longleftarrow \boldsymbol{A}^{22,ij} \text{ component of } \mathcal{S}_6$$

$$E_{\mathfrak{q}}\{(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})\}(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})})^T\} \longleftarrow \boldsymbol{A}^{12,ij} \text{ component of } \mathcal{S}_6$$

$$E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{L1} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{L1})})\}(\boldsymbol{u}_{ij}^{L2} - \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{L2})})^T\} \longleftarrow \boldsymbol{A}^{12,i,j} \text{ component of } \mathcal{S}_6$$

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

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

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

5.3 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 4.3. 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

$$m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u}) = \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i^{\text{L1}} \\ \text{vech}\big(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T\big) \\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T\big) \end{bmatrix} \\ \underset{1\le i\le m}{\text{stack}}\left[\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_{ij}^{\text{L2}} \\ \text{vech}\big(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big) \\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big) \\ \text{vec}\big(\boldsymbol{u}_{ij}^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big) \end{bmatrix}\right] \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2) \to (\boldsymbol{\beta},\boldsymbol{u})}\right\} \tag{24}$$

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 5 points the way to streamlining the fragment updates in the three-level case. Its derivation is given in Section B.11.

Result 5 *The variational message passing updates of the quantities* $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}$, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}$, $1 \le i \le m$, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$, $1 \le i \le m$, $1 \le j \le n_i$, *and the sub-blocks of* $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ *corresponding to (23) with* $\mathfrak{q}$*-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 (see Appendix A.2) with

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{21}^T \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix},$$

$$\boldsymbol{A}_{11} \equiv -2\,\text{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\eta}_{1,2}),$$

$$\boldsymbol{A}_{21} \equiv -\underset{1\le i\le m}{\text{stack}}\Big[\text{vec}^{-1}_{p\times q_1}(\boldsymbol{\eta}_{2,3,i})^T \underset{1\le j\le n_i}{\text{stack}}\{\text{vec}^{-1}_{p\times q_2}(\boldsymbol{\eta}_{3,3,ij})^T\}\Big],$$

$$\boldsymbol{A}_{22} \equiv -2\,\underset{1\le i\le m}{\text{blockdiag}}\begin{bmatrix} \text{vec}^{-1}(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\eta}_{2,2,i}) & \Big[\frac{1}{2}\underset{1\le j\le n_i}{\text{stack}}\{\text{vec}^{-1}_{q_1\times q_2}(\boldsymbol{\eta}_{3,4,ij})^T\}\Big]^T \\ \frac{1}{2}\underset{1\le j\le n_i}{\text{stack}}\{\text{vec}^{-1}_{q_1\times q_2}(\boldsymbol{\eta}_{3,4,ij})^T\} & \underset{1\le j\le n_i}{\text{blockdiag}}\{\text{vec}^{-1}(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\eta}_{3,2,ij})\} \end{bmatrix},$$

and

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

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, according to notation illustrated by (30)–(32),* $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} = \boldsymbol{x}_1$, $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})} = \boldsymbol{A}^{11}$ *and*

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

and

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

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

The message from the likelihood factor to σ^2 has the form as in the two-level case, as given by (19). 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 7. Note its use of the notation defined by (20). Its justification is described in Section B.12.

Algorithm 7 *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: $\boldsymbol{y}_{ij}(o_{ij}\times 1),\ \boldsymbol{X}_{ij}(o_{ij}\times p),\ \boldsymbol{Z}_{ij}^{\text{L1}}(o_{ij}\times q_1),\ \boldsymbol{Z}_{ij}^{\text{L2}}(o_{ij}\times q_2),\ 1\le i\le m,$
$1\le j\le n_i.$

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

Updates:

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

$$\mathcal{S}_7\longleftarrow\textsc{ThreeLevelNaturalToCommonParameters}\Big(p,q_1,q_2,m,\{n_i:1\le i\le m\},\ \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\leftrightarrow(\boldsymbol{\beta},\boldsymbol{u})}\Big)$$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}$ component of $\mathcal{S}_7$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\longleftarrow\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}$ component of $\mathcal{S}_7$

$\boldsymbol{\omega}_{30}\longleftarrow\boldsymbol{0}_p$; $\boldsymbol{\omega}_{31}\longleftarrow\boldsymbol{0}_{\frac{1}{2}p(p+1)}$; $\omega_{32}\longleftarrow 0$

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\longleftarrow\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}$ component of $\mathcal{S}_7$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\longleftarrow\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}$ component of $\mathcal{S}_7$

$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}\longleftarrow E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}$ component of $\mathcal{S}_7$

$\boldsymbol{\omega}_{33i}\longleftarrow\boldsymbol{0}_{q_1}$; $\boldsymbol{\omega}_{34i}\longleftarrow\boldsymbol{0}_{\frac{1}{2}q_1(q_1+1)}$; $\boldsymbol{\omega}_{35i}\longleftarrow\boldsymbol{0}_{p\,q_1}$

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

$\boldsymbol{\omega}_{30}\longleftarrow\boldsymbol{\omega}_{30}+\boldsymbol{X}_{ij}^T\boldsymbol{y}_{ij}$; $\boldsymbol{\omega}_{31}\longleftarrow\boldsymbol{\omega}_{31}-\frac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij})$

$\boldsymbol{\omega}_{33i}\longleftarrow\boldsymbol{\omega}_{33i}+(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{y}_{ij}$; $\boldsymbol{\omega}_{34i}\longleftarrow\boldsymbol{\omega}_{34i}-\frac{1}{2}\boldsymbol{D}_{q_1}^T\text{vec}\big((\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{Z}_{ij}^{\text{L1}}\big)$

$\boldsymbol{\omega}_{35i}\longleftarrow\boldsymbol{\omega}_{35i}-\text{vec}\big(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\text{L1}}\big)$

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\longleftarrow\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ component of $\mathcal{S}_7$

$\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\longleftarrow\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ component of $\mathcal{S}_7$

$E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\longleftarrow E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}$ component of $\mathcal{S}_7$

$E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\longleftarrow$
$E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}$ component of $\mathcal{S}_7$

$\omega_{32}\longleftarrow\omega_{32}-\frac{1}{2}\|\boldsymbol{y}_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}-\boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\|^2$

continued on a subsequent page ...

Algorithm 7 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$
\begin{aligned}
\omega_{32} \longleftarrow \omega_{32} &- \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij}) - \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L1}})\\
&- \tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\boldsymbol{Z}_{ij}^{L2T}\boldsymbol{Z}_{ij}^{L2})\\
&- \mathrm{tr}[\{(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\mathrm{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})^T\}]\\
&- \mathrm{tr}[\{(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\mathrm{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}]\\
&- \mathrm{tr}[\{(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{Z}_{ij}^{\mathrm{L1}}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\mathrm{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})})(\boldsymbol{u}_{ij}^{\mathrm{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})})^T\}]
\end{aligned}
$$

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \mu_{\mathfrak{q}(1/\sigma^2)}\left[\begin{array}{c}
\boldsymbol{\omega}_{30}\\
\boldsymbol{\omega}_{31}\\
\underset{1\le i\le m}{\mathrm{stack}}\left[\begin{array}{c}\boldsymbol{\omega}_{33i}\\ \boldsymbol{\omega}_{34i}\\ \boldsymbol{\omega}_{35i}\end{array}\right]\\
\underset{1\le i\le m}{\mathrm{stack}}\left[\underset{1\le j\le n_i}{\mathrm{stack}}\left[\begin{array}{c}
(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{y}_{ij}\\
-\tfrac{1}{2}\boldsymbol{D}_{q_2}^T\mathrm{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\\
-\mathrm{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\\
-\mathrm{vec}\big((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}}\big)
\end{array}\right]\right]
\end{array}\right]
$$

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,\sigma^2} \longleftarrow \left[\begin{array}{c}
-\tfrac{1}{2}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}o_{ij}\\
\omega_{32}
\end{array}\right]
$$

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

5.4 Streamlined Gaussian Penalization Fragment Updates

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

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

are in the same exponential family. In other words, $m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})$ has the form given by the right-hand side of (24) but with natural parameter vector

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

The fragment's other factor to stochastic node messages are

$$
m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\mathrm{L1}}}(\boldsymbol{\Sigma}^{\mathrm{L1}}) = \exp\left\{\left[\begin{array}{c}\log|\boldsymbol{\Sigma}^{\mathrm{L1}}|\\ \mathrm{vech}\big((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1}\big)\end{array}\right]^T \eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\mathrm{L1}}}\right\}
$$

and

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

Streamlined updating of the three-level Gaussian penalization fragment is aided by Result 6:

Result 6 *The variational message passing updates of the quantities*

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}, \quad 1\le i\le m,$$

and

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}, \quad 1\le i\le m, \quad 1\le j\le n_i,$$

with $\mathfrak{q}$*-density expectations with respect to the normalization of*

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

are expressible as a three-level sparse matrix problem (see Appendix A.2) with

$$\boldsymbol{A} = \begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{21}^T\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix},$$

$$\boldsymbol{A}_{11} \equiv -2\,\mathrm{vec}^{-1}(\boldsymbol{D}_p^{+T}\boldsymbol{\eta}_{1,2}),$$

$$\boldsymbol{A}_{21} \equiv -\underset{1\le i\le m}{\mathrm{stack}}\Big[\mathrm{vec}^{-1}_{p\times q_1}(\boldsymbol{\eta}_{2,3,i})^T\,\underset{1\le j\le n_i}{\mathrm{stack}}\{\mathrm{vec}^{-1}_{p\times q_2}(\boldsymbol{\eta}_{3,3,ij})^T\}\Big],$$

$$\boldsymbol{A}_{22} \equiv -2\,\underset{1\le i\le m}{\mathrm{blockdiag}}\begin{bmatrix}\mathrm{vec}^{-1}(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\eta}_{2,2,i}) & \Big[\frac{1}{2}\underset{1\le j\le n_i}{\mathrm{stack}}\{\mathrm{vec}^{-1}_{q_1\times q_2}(\boldsymbol{\eta}_{3,4,ij})^T\}\Big]^T\\ \frac{1}{2}\underset{1\le j\le n_i}{\mathrm{stack}}\{\mathrm{vec}^{-1}_{q_1\times q_2}(\boldsymbol{\eta}_{3,4,ij})^T\} & \underset{1\le j\le n_i}{\mathrm{blockdiag}}\{\mathrm{vec}^{-1}(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\eta}_{3,2,ij})\}\end{bmatrix},$$

and

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

is the partitioning of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ *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, according to notation illustrated by (30)–(32),*

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

and

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

Algorithm 8 provides the natural parameter vector updates for the three-level Gaussian penalization fragment based on Result 5. Note that natural parameter vectors containing a $\leftrightarrow$ in their subscript, such as $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\mathrm{L1}}}$, are defined by (20).

Algorithm 8 *The inputs, updates and outputs of the matrix algebraic streamlined Gaussian penalization fragment for three-level models. The algorithm description requires more than one page and is continued on a subsequent page.*

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

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

Updates:

$\omega_{36} \longleftarrow$ first entry of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\mathrm{L1}}}$

$\boldsymbol{\omega}_{37} \longleftarrow$ remaining entries of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\mathrm{L1}}}$

$\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L1}})^{-1})} \longleftarrow \big\{\omega_{36} + \frac{1}{2}(q_1+1)\big\}\{\mathrm{vec}^{-1}\big(\boldsymbol{D}_{q_1}^{+T}\boldsymbol{\omega}_{37}\big)\}^{-1}$

$\omega_{38} \longleftarrow$ first entry of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\mathrm{L2}}}$

$\boldsymbol{\omega}_{39} \longleftarrow$ remaining entries of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\mathrm{L1}},\boldsymbol{\Sigma}^{\mathrm{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\mathrm{L2}}}$

$\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\mathrm{L2}})^{-1})} \longleftarrow \big\{\omega_{38} + \frac{1}{2}(q_2+1)\big\}\{\mathrm{vec}^{-1}\big(\boldsymbol{D}_{q_2}^{+T}\boldsymbol{\omega}_{39}\big)\}^{-1}$

$\mathcal{S}_8 \longleftarrow$ ThreeLevelNaturalToCommonParameters
$\Big(p, q_1, q_2, m, \{n_i : 1 \le i \le m\}, \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}\Big)$

$\boldsymbol{\omega}_{40} \longleftarrow \boldsymbol{0}_{\frac{1}{2}q_1(q_1+1)}$; $\boldsymbol{\omega}_{41} \longleftarrow \boldsymbol{0}_{\frac{1}{2}q_2(q_2+1)}$

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})} \longleftarrow \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}$ component of $\mathcal{S}_8$

$\boldsymbol{\omega}_{40} \longleftarrow \boldsymbol{\omega}_{40} - \frac{1}{2}\boldsymbol{D}_{q_1}^T\mathrm{vec}\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\Big)$

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

$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})} \longleftarrow \boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$ component of $\mathcal{S}_8$; $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})} \longleftarrow \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}$ component of $\mathcal{S}_8$

$\boldsymbol{\omega}_{41} \longleftarrow \boldsymbol{\omega}_{41} - \frac{1}{2}\boldsymbol{D}_{q_2}^T\mathrm{vec}\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\Big)$

continued on a subsequent page ...

Algorithm 8 continued. *This is a continuation of the description of this algorithm that commences on a preceding page.*

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \mathbf{0}_{q_1} \\ -\tfrac{1}{2}\boldsymbol{D}_{q_1}^T\text{vec}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})}\big) \\ \mathbf{0}_{pq_1} \end{bmatrix} \\ \underset{1\le i\le m}{\text{stack}}\left[\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix} \mathbf{0}_{q_2} \\ -\tfrac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}\big(\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}\big) \\ \mathbf{0}_{pq_2} \\ \mathbf{0}_{q_1q_2} \end{bmatrix}\right] \end{bmatrix}$$

$$\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L1}}} \longleftarrow \begin{bmatrix} -\tfrac{1}{2}m \\ \boldsymbol{\omega}_{40} \end{bmatrix} \quad ; \quad \eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L2}}} \longleftarrow \begin{bmatrix} -\tfrac{1}{2}\sum_{i=1}^{m} n_i \\ \boldsymbol{\omega}_{41} \end{bmatrix}$$

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

5.5 $\mathfrak{q}$-Density Determination After Variational Message Passing Convergence

The advice given in Section 4.5 for the two-level case extends straightforwardly to the three-level case. The main change is that the steps that we need to first carry out are:

$$\begin{aligned} &\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})} \\ &\mathcal{S}_B \longleftarrow \textsc{ThreeLevelNaturalToCommonParameters} \\ &\qquad\Big(p, q_1, q_2, m, \{n_i : 1 \le i \le m\}, \boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}\Big). \end{aligned}$$

6. Computational Complexity and Timing Results

Table 1 summarizes and compares the large sample computational complexities of streamlined mean field variational Bayes Algorithms 1 and 5 and the naïve implementation alternative. To aid digestibility, in Table 1 we are imposing the following balanced designs restrictions: $n_i = n$ and $o_{ij} = o$ for all values of the indices. The values of m, n and o are assumed to be diverging whilst p, q, q_1, q_2 and the numbers of mean field variational Bayes iterations are held fixed. The entries of Table 1 are justified by results concerning the number of floating point operations for matrix multiplications and QR decompositions given in, for example, Sections 1.2.4 and 5.5.9 of Golub and Loan (1989). We see from Table

1 that the floating point operation counts of Algorithms 1 and 5 are linear in the number of observations and these streamlined algorithms offer quadratic improvements over naïve implementation.

level	naïve	streamlined	naïve/streamlined
two-level	$O(m^3n)$	$O(mn)$	$O(m^2)$
three-level	$O(m^3n^3o)$	$O(mno)$	$O(m^2n^2)$

Table 1: *The order of magnitudes of the number of floating point operations for Algorithms 1 and 5 and naïve implementation. The ratio of naïve to streamlined computation is also given. The designs are assumed to be balanced and $m, n, o \to \infty$ whilst p, q, q_1, q_2 and the numbers of mean field variational Bayes iterations are fixed.*

To assess finite sample performance, we obtained timing results for simulated data according to a version of model (7) for which both the fixed effects and random effects had dimension 2, 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

$$\boldsymbol{\beta}_{\text{true}} = \begin{bmatrix} 0.58 \\ 1.98 \end{bmatrix}, \qquad \sigma^2_{\text{true}} = 0.1 \qquad \text{and} \qquad \boldsymbol{\Sigma}_{\text{true}} = \begin{bmatrix} 2.58 & 0.22 \\ 0.22 & 1.73 \end{bmatrix}$$

and, throughout the study, the n_i values were generated uniformly on the set $\{30, \ldots, 60\}$. 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
1,600	427.00 (3.1000)	0.183 (0.00741)	2340.0

Table 2: *Median (median absolute deviation) of elapsed computing times in seconds for fitting model (7) naïvely versus with streamlining via Algorithm 1. 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 1 with its naïve counterpart for which (10) 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 in the naïve computational times our summary of this first phase, given in Table 2, uses medians and median absolute deviations. As the number of groups increases into the several 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 1. We used the geometric progression $m \in \{400, 1200, 3600, 10800, 32400\}$ and another 100 replications. Table 3 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 (0.0122)	0.2400 (0.0343)	0.7140 (0.0806)	2.30 (0.270)	6.980 (0.857)

Table 3: *Average (standard deviation) of elapsed computing times in seconds for fitting model (7) with streamlining via Algorithm 1.*

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

As a by-product of our timing studies we also recorded the empirical coverage percentages for credible intervals with an advertized coverage of 95%. The results are given in Table 4 and based on 1,000 replications. Apart from σ, the parameters in Table 4 are sub-components of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ according to

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Taking into account the margins of error in percentage estimates based on 1,000 replications, the empirical coverages are seen to be in keeping with the 95% advertized level.

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 the 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 interpretation. Instead we consider an illustrative Bayesian mixed model, with two-level random effects, and compare streamlined

parameter	$m = 100$	$m = 200$	$m = 400$	$m = 800$	$m = 1,600$
β_0	96.2	95.0	95.6	94.7	95.3
β_1	94.8	95.2	94.5	95.4	93.5
σ	95.1	94.0	95.3	95.1	94.6
σ_1	93.8	93.6	95.1	95.2	95.3
σ_2	94.3	94.3	93.9	95.5	95.3
ρ	93.9	95.9	95.1	95.0	93.8

Table 4: *Empirical coverage percentages for advertized 95% credible intervals produced by Algorithm 1 for the simulation study described in the text. The empirical coverage percentages are based on* 1,000 *replications.*

mean field variational Bayes and Markov chain Monte Carlo fits. Specifically, we consider the model

$$
\begin{aligned}
y_{ij}|\beta_0,\ldots,\beta_7,\sigma^2 &\stackrel{\text{ind.}}{\sim} N\Big(\beta_0 + u_{i0} + (\beta_1 + u_{i1})x_{1ij} + (\beta_1 + u_{i2})x_{2ij}^2 \\
&\qquad\qquad + \beta_3\, x_{3ij} + \ldots + \beta_7\, x_{7ij}, \sigma^2\Big), \\
\left.\begin{bmatrix} u_{i0} \\ u_{i1} \\ u_{i2} \end{bmatrix}\right|\boldsymbol{\Sigma} &\stackrel{\text{ind.}}{\sim} N(\mathbf{0},\boldsymbol{\Sigma}),\ \text{for } 1 \le i \le 44,708 \text{ and } 1 \le j \le n_i
\end{aligned}
\tag{25}
$$

with priors

$$
\begin{gathered}
\beta_0,\ldots,\beta_7 \stackrel{\text{ind.}}{\sim} N(0,10^{10}), \quad \sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(1,1/a_{\sigma^2}), \\
a_{\sigma^2} \sim \text{Inverse-}\chi^2(1,10^{-10}), \quad \boldsymbol{\Sigma}|\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}\big(G_{\text{full}}, 6, \boldsymbol{A}_{\boldsymbol{\Sigma}}^{-1}\big), \\
\boldsymbol{A}_{\boldsymbol{\Sigma}} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, 2\times 10^{-10}\boldsymbol{I}_3)
\end{gathered}
\tag{26}
$$

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.

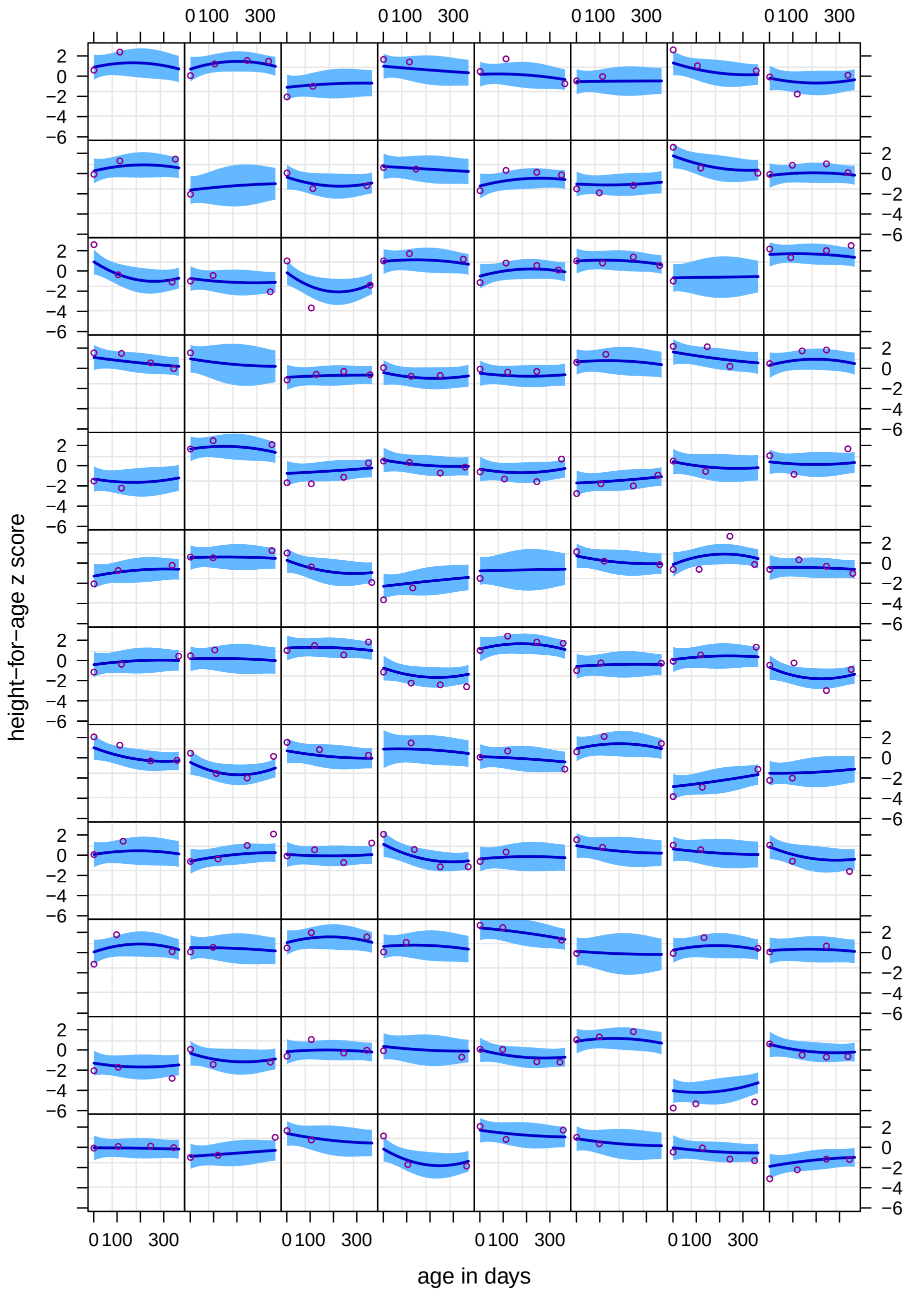

Figure 7: *Fitted random quadratics for 96 randomly chosen infants from the streamlined mean field variational Bayes analysis of data from the Collaborative Perinatal Project for infants in the first year of life. The curves correspond to slices of the fitted surface according to the model defined by (25) and (26) with each of the other predictors set to its average value. The light blue shading corresponds to pointwise 95% credible intervals.*

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 (25) with priors (26) 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 (25) 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 apparent from Figure 7 that the curvature in the age effects warrants the extension to random quadratics.

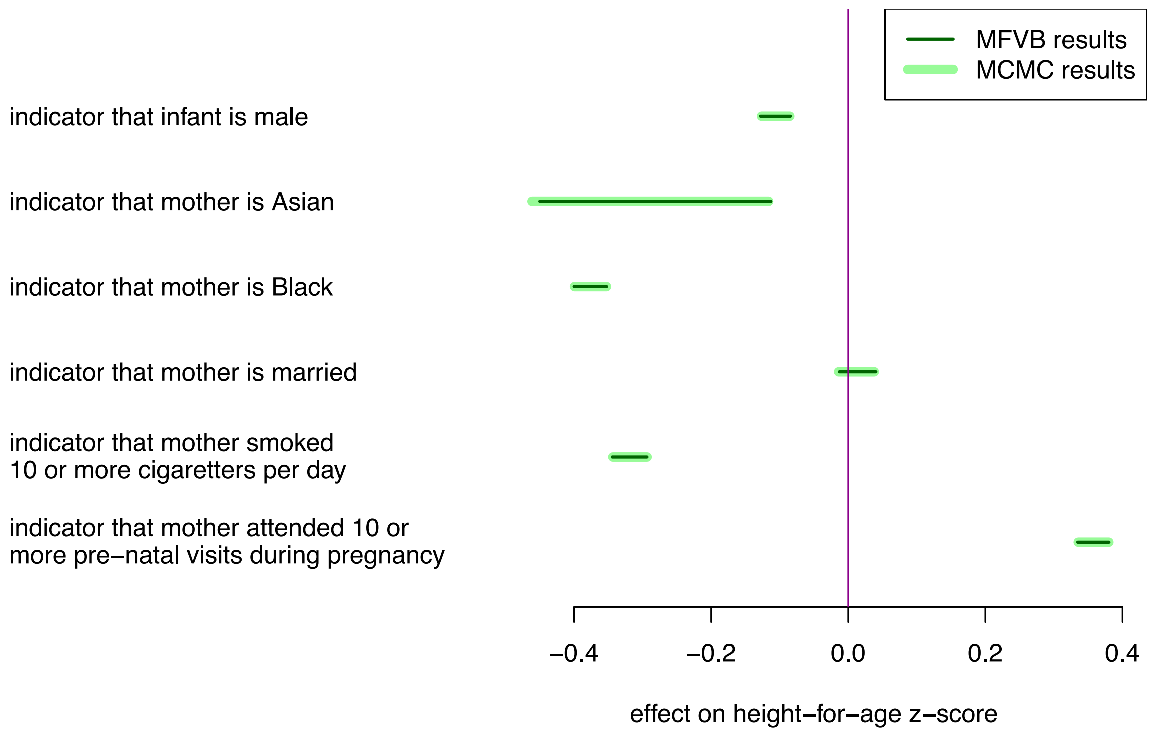

Figure 8: *Approximate 95% credible intervals for $\beta_3, \ldots, \beta_7$ for two approximate Bayesian inference fits to the model defined by (25) and (26) for the data from the Collaborative Perinatal Project for infants in the first year of life. The thin dark green line segments display credible 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.*

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 rstan (Stan Development Team, 2020), the R (R Core Team, 2020) interface to the Stan language, 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. As explained in Section 3.1 of Menictas and Wand (2013), mean field variational Bayes tends to be very accurate for Gaussian response models of the type being used in this example and the mild product restriction (8). However, such high accuracy is not manifest in general. Ignorance of important posterior dependencies via mean field restrictions often lead to credible intervals being too small (e.g. Wang and Titterington, 2005). In Figure 8 there are pronounced negative effects due to ethnicity and maternal smoking and a pronounced positive effect due to pre-natal care.

Even though streamlined mean field variational Bayes and Markov chain Monte Carlo deliver similar inference for this example, the former is significantly faster. However, it is difficult to quantify the speed gains scientifically due to factors such as stopping criteria, implementation language and quality of the chains. For the Figure 8 fits, using the `MacBook Air` laptop described in Section 6 the Markov chain Monte Carlo fits required about 36 hours whilst the streamlined variational results took just 24 seconds. However, this comparison is based on a *convenient* version of Markov chain Monte Carlo in which all the user has to do is specify the model and let the Stan Bayesian inference engine do the work. This convenience comes at the cost that general purpose Bayesian inference engines tend to be slower than Markov chain Monte implementations for specific models. For the model and priors given by (25) and (26) Gibbs sampling involves standard distributions and can be streamlined by sampling from the fixed effects vector and then looping through the random effect vectors for each infant. After carrying out the requisite algebra, and programming streamlined Gibbs sampling in R, we found that Markov chain Monte Carlo fitting with the same chain sizes and laptop required about 3.5 hours. This is about 10 times faster than Stan, but took a lot longer to code. Lastly, we implemented streamlined Gibbs sampling using the low-level C++ language with the aid of the R packages Rcpp (Eddelbeuttel et al., 2020a), RcppArmadillo (Eddelbeuttel et al., 2020b) and RcppDist (Duck-Mayr, 2018). The coding time required by the authors for this C++ implementation was much longer than using Stan, but it resulted in a fitting time of just 4.9 minutes. Compared with Stan, the quality of the chains produced by these streamlined Gibbs sampling implementations is not as high and larger warmup and kept sample sizes may be warranted in practice.

Table 5 summarizes all of the timings for this example. It shows that, depending on how Markov chain Monte Carlo is implemented, Bayesian linear mixed model analysis of the Collaborative Perinatal Project data is between several thousand times and a dozen times slower than streamlined variational inference.

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 and Wand (2016). Another is non-Gaussian likelihood fragments. Group specific curve models (e.g. Durban et al., 2005) also

approach	computing time	MCMC/(streamlined MFVB)
MCMC via rstan	36 hours	5,400
MCMC via R code	3.5 hours	514
MCMC via C++ code	4.9 minutes	12.3
streamlined MFVB	24 seconds	—

Table 5: *Computing times for four different approaches to approximate Bayesian fitting of (25) to the Collaborative Perinatal Project data. The first three approaches are Markov chain Monte Carlo (MCMC) with a warmup of length 1,000 and then 1,000 retained samples. The last approach is streamlined mean field variational Bayes (MFVB) with 100 iterations. The ratios of the MCMC computing times to that of streamlined MFVB are also shown.*

lend themselves to streamlining via the SolveTwoLevelSparseLeastSquares and SolveThreeLevelSparseLeastSquares algorithms and Menictas et al. (2020) 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.

Acknowledgments

We are grateful to Hon Hwang, Robert Kohn, Cathy Lee, Luca Maestrini, Chris Oates, Louise Ryan, Christian Tezza, the editor and two reviewers for their valuable contributions. This research was supported by Australian Research Council Discovery Project DP180100597 and aided by the Knowledge integration project within the Bill and Melinda Gates Foundation.

Appendix A. Multilevel Sparse Matrix Problem Algorithms

Algorithms 1–8 rely on four fundamental matrix algebraic algorithms that solve the two-level and three-level versions of *multilevel sparse matrix problems*. This class of problems are defined in Nolan and Wand (2020). These four algorithms:

SolveTwoLevelSparseMatrix	Algorithm A.1
SolveTwoLevelSparseLeastSquares	Algorithm A.2
SolveThreeLevelSparseMatrix	Algorithm A.3
SolveThreeLevelSparseLeastSquares	Algorithm A.4

and their underpinnings are presented in this appendix.

A.1 Two-Level Sparse Matrix Algorithms

Two-level sparse matrix problems are described in Section 2 of Nolan and Wand (2020). The notation used there is also used in this section. Here we present two algorithms, named

SolveTwoLevelSparseMatrix and 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

$$\boldsymbol{A} \equiv \left[\begin{array}{c|c|c|c|c} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \cdots & \boldsymbol{A}_{12,m} \\ \hline \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \hline \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \cdots & \boldsymbol{O} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{A}_{12,m}^T & \boldsymbol{O} & \boldsymbol{O} & \cdots & \boldsymbol{A}_{22,m} \end{array}\right], \quad \boldsymbol{a} \equiv \left[\begin{array}{c} \boldsymbol{a}_1 \\ \hline \boldsymbol{a}_{2,1} \\ \hline \boldsymbol{a}_{2,2} \\ \hline \vdots \\ \hline \boldsymbol{a}_{2,m} \end{array}\right] \quad \text{and} \quad \boldsymbol{x} \equiv \left[\begin{array}{c} \boldsymbol{x}_1 \\ \hline \boldsymbol{x}_{2,1} \\ \hline \boldsymbol{x}_{2,2} \\ \hline \vdots \\ \hline \boldsymbol{x}_{2,m} \end{array}\right] \tag{27}$$

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

$$\boldsymbol{A}^{-1} \equiv \left[\begin{array}{c|c|c|c|c} \boldsymbol{A}^{11} & \boldsymbol{A}^{12,1} & \boldsymbol{A}^{12,2} & \cdots & \boldsymbol{A}^{12,m} \\ \hline \boldsymbol{A}^{12,1\,T} & \boldsymbol{A}^{22,1} & \times & \cdots & \times \\ \hline \boldsymbol{A}^{12,2\,T} & \times & \boldsymbol{A}^{22,2} & \cdots & \times \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{A}^{12,m\,T} & \times & \times & \cdots & \boldsymbol{A}^{22,m} \end{array}\right]. \tag{28}$$

As will be elaborated upon later, the blocks represented by the $\times$ symbol are not of interest. SolveTwoLevelSparseMatrix is listed as Algorithm A.1 and is justified by Theorem 2.2 of Nolan and Wand (2020).

The SolveTwoLevelSparseLeastSquares algorithm arises in the special case 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 the matrix $\boldsymbol{B}$ and vector $\boldsymbol{b}$ have the generic forms

$$\boldsymbol{B} \equiv \left[\begin{array}{c|c|c|c|c} \boldsymbol{B}_1 & \dot{\boldsymbol{B}}_1 & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \hline \boldsymbol{B}_2 & \boldsymbol{O} & \dot{\boldsymbol{B}}_2 & \cdots & \boldsymbol{O} \\ \hline \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{B}_m & \boldsymbol{O} & \boldsymbol{O} & \cdots & \dot{\boldsymbol{B}}_m \end{array}\right] \quad \text{and} \quad \boldsymbol{b} \equiv \left[\begin{array}{c} \boldsymbol{b}_1 \\ \hline \boldsymbol{b}_2 \\ \hline \vdots \\ \hline \boldsymbol{b}_m \end{array}\right]. \tag{29}$$

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

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

Algorithm A.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 (27) and (28).*

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

$\boldsymbol{\omega}_{42} \longleftarrow \boldsymbol{a}_1 \quad ; \quad \boldsymbol{\Omega}_{43} \longleftarrow \boldsymbol{A}_{11}$

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

$\quad\boldsymbol{\omega}_{42} \longleftarrow \boldsymbol{\omega}_{42} - \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{a}_{2,i} \quad ; \quad \boldsymbol{\Omega}_{43} \longleftarrow \boldsymbol{\Omega}_{43} - \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^T$

$\boldsymbol{A}^{11} \longleftarrow \boldsymbol{\Omega}_{43}^{-1} \quad ; \quad \boldsymbol{x}_1 \longleftarrow \boldsymbol{A}^{11}\boldsymbol{\omega}_{42}$

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

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

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

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

As demonstrated in Section 4, these forms arise in two-level random effects models. Theorem 2.3 of Nolan and Wand (2020) 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 A.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.

Note that in Algorithm A.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.

A.2 Three-Level Sparse Matrix Algorithms

Extension to the three-level situation is described in Section 3 of Nolan and Wand (2020). Theorems 3.2 and 3.3 given there lead to the algorithms

SolveThreeLevelSparseMatrix and SolveThreeLevelSparseLeastSquares

which facilitate streamlining variational inference for three-level models.

Algorithm A.2 *The* 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 (27), (28) and (29).*

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

$\boldsymbol{\omega}_{44}\longleftarrow$ NULL ; $\boldsymbol{\Omega}_{45}\longleftarrow$ NULL

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

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

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

$\boldsymbol{c}_{1i}\longleftarrow$ first q rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{c}_{2i}\longleftarrow$ remaining rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{\omega}_{44}\longleftarrow\begin{bmatrix}\boldsymbol{\omega}_{44}\\ \boldsymbol{c}_{2i}\end{bmatrix}$

$\boldsymbol{C}_{1i}\longleftarrow$ first q rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{C}_{2i}\longleftarrow$ remaining rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{\Omega}_{45}\longleftarrow\begin{bmatrix}\boldsymbol{\Omega}_{45}\\ \boldsymbol{C}_{2i}\end{bmatrix}$

Decompose $\boldsymbol{\Omega}_{45}=\boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{bmatrix}$ such that $\boldsymbol{Q}^{-1}=\boldsymbol{Q}^T$ and $\boldsymbol{R}$ is upper-triangular.

$\boldsymbol{c}\longleftarrow$ first p rows of $\boldsymbol{Q}^T\boldsymbol{\omega}_{44}$; $\boldsymbol{x}_1\longleftarrow\boldsymbol{R}^{-1}\boldsymbol{c}$; $\boldsymbol{A}^{11}\longleftarrow\boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$

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

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

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

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

An illustrative three-level sparse matrix is:

$$
\boldsymbol{A}=\left[\begin{array}{c|c|c|c|c|c|c|c}
\boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,11} & \boldsymbol{A}_{12,12} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,21} & \boldsymbol{A}_{12,22} & \boldsymbol{A}_{12,23}\\ \hline
\boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{A}_{12,1,1} & \boldsymbol{A}_{12,1,2} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O}\\ \hline
\boldsymbol{A}_{12,11}^T & \boldsymbol{A}_{12,1,1}^T & \boldsymbol{A}_{22,11} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O}\\ \hline
\boldsymbol{A}_{12,12}^T & \boldsymbol{A}_{12,1,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,12} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O}\\ \hline
\boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{A}_{12,2,1} & \boldsymbol{A}_{12,2,2} & \boldsymbol{A}_{12,2,3}\\ \hline
\boldsymbol{A}_{12,21}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{12,2,1}^T & \boldsymbol{A}_{22,21} & \boldsymbol{O} & \boldsymbol{O}\\ \hline
\boldsymbol{A}_{12,22}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{12,2,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,22} & \boldsymbol{O}\\ \hline
\boldsymbol{A}_{12,23}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{12,2,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,23}
\end{array}\right] \tag{30}
$$

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 $\left\{\left[\begin{array}{c|c|c|c} \boldsymbol{A}_{12,i} & \boldsymbol{A}_{12,ij} & \ldots & \boldsymbol{A}_{12,in_i} \end{array}\right] : 1 \leq i \leq m\right\}$, 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 (27). 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:

$$\boldsymbol{a} \equiv \left[\begin{array}{c} \boldsymbol{a}_1 \\ \hline \boldsymbol{a}_{2,1} \\ \hline \boldsymbol{a}_{2,11} \\ \hline \boldsymbol{a}_{2,12} \\ \hline \boldsymbol{a}_{2,2} \\ \hline \boldsymbol{a}_{2,21} \\ \hline \boldsymbol{a}_{2,22} \\ \hline \boldsymbol{a}_{2,23} \end{array}\right] \quad \text{and} \quad \boldsymbol{x} \equiv \left[\begin{array}{c} \boldsymbol{x}_1 \\ \hline \boldsymbol{x}_{2,1} \\ \hline \boldsymbol{x}_{2,11} \\ \hline \boldsymbol{x}_{2,12} \\ \hline \boldsymbol{x}_{2,2} \\ \hline \boldsymbol{x}_{2,21} \\ \hline \boldsymbol{x}_{2,22} \\ \hline \boldsymbol{x}_{2,23} \end{array}\right]. \tag{31}$$

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

$$
\boldsymbol{A}^{-1}=\left[\begin{array}{c|c|c|c|c|c|c|c}
\boldsymbol{A}^{11} & \boldsymbol{A}^{12,1} & \boldsymbol{A}^{12,11} & \boldsymbol{A}^{12,12} & \boldsymbol{A}^{12,2} & \boldsymbol{A}^{12,21} & \boldsymbol{A}^{12,22} & \boldsymbol{A}^{12,23} \\
\hline
\boldsymbol{A}^{12,1\,T} & \boldsymbol{A}^{22,1} & \boldsymbol{A}^{12,1,1} & \boldsymbol{A}^{12,1,2} & \times & \times & \times & \times \\
\hline
\boldsymbol{A}^{12,11\,T} & \boldsymbol{A}^{12,1,1\,T} & \boldsymbol{A}^{22,11} & \times & \times & \times & \times & \times \\
\hline
\boldsymbol{A}^{12,12\,T} & \boldsymbol{A}^{12,1,2\,T} & \times & \boldsymbol{A}^{22,12} & \times & \times & \times & \times \\
\hline
\boldsymbol{A}^{12,2\,T} & \times & \times & \times & \boldsymbol{A}^{22,2} & \boldsymbol{A}^{12,2,1} & \boldsymbol{A}^{12,2,2} & \boldsymbol{A}^{12,2,3} \\
\hline
\boldsymbol{A}^{12,21\,T} & \times & \times & \times & \boldsymbol{A}^{12,2,1\,T} & \boldsymbol{A}^{22,21} & \times & \times \\
\hline
\boldsymbol{A}^{12,22\,T} & \times & \times & \times & \boldsymbol{A}^{12,2,2\,T} & \times & \boldsymbol{A}^{22,22} & \times \\
\hline
\boldsymbol{A}^{12,23\,T} & \times & \times & \times & \boldsymbol{A}^{12,2,3\,T} & \times & \times & \boldsymbol{A}^{22,23}
\end{array}\right] \tag{32}
$$

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

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

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 $\|\boldsymbol{b}-\boldsymbol{B}\boldsymbol{x}\|^2 \equiv (\boldsymbol{b}-\boldsymbol{B}\boldsymbol{x})^T(\boldsymbol{b}-\boldsymbol{B}\boldsymbol{x})$ 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

$$
\boldsymbol{B}\equiv\left[\begin{array}{c|c|c|c|c|c|c|c}
\boldsymbol{B}_{11} & \dot{\boldsymbol{B}}_{11} & \ddot{\boldsymbol{B}}_{11} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\
\hline
\boldsymbol{B}_{12} & \dot{\boldsymbol{B}}_{12} & \boldsymbol{O} & \ddot{\boldsymbol{B}}_{12} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\
\hline
\boldsymbol{B}_{21} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dot{\boldsymbol{B}}_{21} & \ddot{\boldsymbol{B}}_{21} & \boldsymbol{O} & \boldsymbol{O} \\
\hline
\boldsymbol{B}_{22} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dot{\boldsymbol{B}}_{22} & \boldsymbol{O} & \ddot{\boldsymbol{B}}_{22} & \boldsymbol{O} \\
\hline
\boldsymbol{B}_{23} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \dot{\boldsymbol{B}}_{23} & \boldsymbol{O} & \boldsymbol{O} & \ddot{\boldsymbol{B}}_{23}
\end{array}\right] \quad \text{and} \quad \boldsymbol{b}\equiv\left[\begin{array}{c}
\boldsymbol{b}_{11} \\ \hline \boldsymbol{b}_{12} \\ \hline \boldsymbol{b}_{21} \\ \hline \boldsymbol{b}_{22} \\ \hline \boldsymbol{b}_{23}
\end{array}\right]. \tag{33}
$$

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:

$$
\boldsymbol{b}_{ij} \text{ is } \tilde{o}_{ij}\times 1, \quad \boldsymbol{B}_{ij} \text{ is } \tilde{o}_{ij}\times p, \quad \dot{\boldsymbol{B}}_{ij} \text{ is } \tilde{o}_{ij}\times q_1, \quad \text{and} \quad \ddot{\boldsymbol{B}}_{ij} \text{ is } \tilde{o}_{ij}\ \times q_2. \tag{34}
$$

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. The general forms of $\boldsymbol{B}$ and $\boldsymbol{b}$ in the three-level case are

$$
\begin{array}{l}
\boldsymbol{B}\equiv\left[\underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le j\le n_i}{\text{stack}}(\boldsymbol{B}_{ij})\right\}\ \Big|\underset{1\le i\le m}{\text{blockdiag}}\left\{\left[\underset{1\le j\le n_i}{\text{stack}}(\dot{\boldsymbol{B}}_{ij})\ \Big|\underset{1\le j\le n_i}{\text{blockdiag}}(\ddot{\boldsymbol{B}}_{ij})\right]\right\}\right] \\
\text{and} \quad \boldsymbol{b}\equiv\underset{1\le i\le m}{\text{stack}}\left\{\underset{1\le j\le n_i}{\text{stack}}(\boldsymbol{b}_{ij})\right\}.
\end{array} \tag{35}
$$

Algorithm A.4 provides a QR decomposition-based solution to the three-level sparse matrix least squares problems when the inputs are the matrices listed in (34).

Algorithm A.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 (30), (31) and (32).*

Input: $\Big(\boldsymbol{a}_1(p\times 1), \boldsymbol{A}_{11}(p\times p), \big\{\big(\boldsymbol{a}_{2,i}(q_1\times 1), \boldsymbol{A}_{22,i}(q_1\times q_1), \boldsymbol{A}_{12,i}(p\times q_1) :\ 1\le i\le m\big\},$
$\big\{\boldsymbol{a}_{2,ij}(q_2\times 1), \boldsymbol{A}_{22,ij}(q_2\times q_2), \boldsymbol{A}_{12,ij}(p\times q_2), \boldsymbol{A}_{12,\,i,\,j}(q_1\times q_2)\big) :\ 1\le i\le m,\ 1\le j\le n_i\big\}\Big).$

$\boldsymbol{\omega}_{46} \longleftarrow \boldsymbol{a}_1 \quad ; \quad \boldsymbol{\Omega}_{47} \longleftarrow \boldsymbol{A}_{11}$

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

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

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

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

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

$\quad\quad \boldsymbol{\omega}_{46} \longleftarrow \boldsymbol{\omega}_{46} - \boldsymbol{A}_{12,ij}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{a}_{2,ij} \quad ; \quad \boldsymbol{\Omega}_{47} \longleftarrow \boldsymbol{\Omega}_{47} - \boldsymbol{A}_{12,ij}\boldsymbol{A}_{22,ij}^{-1}\boldsymbol{A}_{12,ij}^T$

$\quad \boldsymbol{\omega}_{46} \longleftarrow \boldsymbol{\omega}_{46} - \boldsymbol{H}_{12,i}\boldsymbol{H}_{22,i}^{-1}\boldsymbol{h}_{2,i} \quad ; \quad \boldsymbol{\Omega}_{47} \longleftarrow \boldsymbol{\Omega}_{47} - \boldsymbol{H}_{12,i}\boldsymbol{H}_{22,i}^{-1}\boldsymbol{H}_{12,i}^T$

$\boldsymbol{A}^{11} \longleftarrow \boldsymbol{\Omega}_{47}^{-1} \quad ; \quad \boldsymbol{x}_1 \longleftarrow \boldsymbol{A}^{11}\boldsymbol{\omega}_{46}$

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

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

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

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

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

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

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

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

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

Appendix B. Appendix B. Derivations

B.1 Derivation of Result 1

It is straightforward to verify that the $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ updates, given at (10), may be written as

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

Algorithm A.4 *The* 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 (28). The algorithm description requires more than one page and is continued on a subsequent page.*

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

$\boldsymbol{\omega}_{48}\longleftarrow \text{NULL}\quad;\quad \boldsymbol{\Omega}_{49}\longleftarrow\text{NULL}$

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

$\boldsymbol{\omega}_{50}\longleftarrow \text{NULL}\quad;\quad \boldsymbol{\Omega}_{51}\longleftarrow\text{NULL}\quad;\quad \boldsymbol{\Omega}_{52}\longleftarrow\text{NULL}$

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

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.

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

$\boldsymbol{d}_{1ij}\longleftarrow$ 1st q_2 rows of $\boldsymbol{d}_{0ij}$; $\boldsymbol{d}_{2ij}\longleftarrow$ remaining rows of $\boldsymbol{d}_{0ij}$; $\boldsymbol{\omega}_{50}\longleftarrow\begin{bmatrix}\boldsymbol{\omega}_{50}\\ \boldsymbol{d}_{2ij}\end{bmatrix}$

$\boldsymbol{D}_{1ij}\longleftarrow$ 1st q_2 rows of $\boldsymbol{D}_{0ij}$; $\boldsymbol{D}_{2ij}\longleftarrow$ remaining rows of $\boldsymbol{D}_{0ij}$

$\boldsymbol{\Omega}_{51}\longleftarrow\begin{bmatrix}\boldsymbol{\Omega}_{51}\\ \boldsymbol{D}_{2ij}\end{bmatrix}$

$\dot{\boldsymbol{D}}_{1ij}\longleftarrow$ 1st q_2 rows of $\dot{\boldsymbol{D}}_{0ij}$; $\dot{\boldsymbol{D}}_{2ij}\longleftarrow$ remaining rows of $\dot{\boldsymbol{D}}_{0ij}$

$\boldsymbol{\Omega}_{52}\longleftarrow\begin{bmatrix}\boldsymbol{\Omega}_{52}\\ \dot{\boldsymbol{D}}_{2ij}\end{bmatrix}$

Decompose $\boldsymbol{\Omega}_{52}=\boldsymbol{Q}_i\begin{bmatrix}\boldsymbol{R}_i\\ \boldsymbol{O}\end{bmatrix}$ such that $\boldsymbol{Q}_i^{-1}=\boldsymbol{Q}_i^T$ and $\boldsymbol{R}_i$ is upper-triangular.

$\boldsymbol{c}_{0i}\longleftarrow \boldsymbol{Q}_i^T\boldsymbol{\omega}_{50}\quad;\quad \boldsymbol{C}_{0i}\longleftarrow \boldsymbol{Q}_i^T\boldsymbol{\Omega}_{51}$

$\boldsymbol{c}_{1i}\longleftarrow$ 1st q_1 rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{c}_{2i}\longleftarrow$ remaining rows of $\boldsymbol{c}_{0i}$; $\boldsymbol{\omega}_{48}\longleftarrow\begin{bmatrix}\boldsymbol{\omega}_{48}\\ \boldsymbol{c}_{2i}\end{bmatrix}$

$\boldsymbol{C}_{1i}\longleftarrow$ 1st q_1 rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{C}_{2i}\longleftarrow$ remaining rows of $\boldsymbol{C}_{0i}$; $\boldsymbol{\Omega}_{49}\longleftarrow\begin{bmatrix}\boldsymbol{\Omega}_{49}\\ \boldsymbol{C}_{2i}\end{bmatrix}$

Decompose $\boldsymbol{\Omega}_{49}=\boldsymbol{Q}\begin{bmatrix}\boldsymbol{R}\\ \boldsymbol{O}\end{bmatrix}$ so that $\boldsymbol{Q}^{-1}=\boldsymbol{Q}^T$ and $\boldsymbol{R}$ is upper-triangular.

$\boldsymbol{c}\longleftarrow$ first p rows of $\boldsymbol{Q}^T\boldsymbol{\omega}_{48}$; $\boldsymbol{x}_1\longleftarrow \boldsymbol{R}^{-1}\boldsymbol{c}$; $\boldsymbol{A}^{11}\longleftarrow \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$

continued on a subsequent page ...

where $\boldsymbol{B}$ and $\boldsymbol{b}$ have the forms (29) with

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

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

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

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

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

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

$\boldsymbol{x}_{2,ij} \leftarrow \boldsymbol{R}_{ij}^{-1}(\boldsymbol{d}_{1ij} - \boldsymbol{D}_{1ij}\boldsymbol{x}_1 - \overset{\bullet}{\boldsymbol{D}}_{1ij}\boldsymbol{x}_{2,i})$

$\boldsymbol{A}^{12,ij} \leftarrow -\Big\{\boldsymbol{R}_{ij}^{-1}(\boldsymbol{D}_{1ij}\boldsymbol{A}^{11} + \overset{\bullet}{\boldsymbol{D}}_{1ij}\boldsymbol{A}^{12,i\,T})\Big\}^T$

$\boldsymbol{A}^{12,\,i,\,j} \leftarrow -\Big\{\boldsymbol{R}_{ij}^{-1}(\boldsymbol{D}_{1ij}\boldsymbol{A}^{12,i} + \overset{\bullet}{\boldsymbol{D}}_{1ij}\boldsymbol{A}^{22,i})\Big\}^T$

$\boldsymbol{A}^{22,ij} \leftarrow \boldsymbol{R}_{ij}^{-1}\big(\boldsymbol{R}_{ij}^{-T} - \boldsymbol{D}_{1ij}\boldsymbol{A}^{12,ij} - \overset{\bullet}{\boldsymbol{D}}_{1ij}\boldsymbol{A}^{12,\,i,\,j}\big)$

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

B.2 Derivation of Algorithm 1

We first provide expressions for the $\mathfrak{q}$-densities for mean field variational Bayesian inference for the parameters in (7), with product density restriction (8). Arguments analogous to those given in, for example, Appendix C of Wand and Ormerod (2011) lead to:

$$\mathfrak{q}(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is a } N(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}) \text{ density function}$$

where

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

with $\boldsymbol{R}_{\text{MFVB}}$, $\boldsymbol{D}_{\text{MFVB}}$ and $\boldsymbol{o}_{\text{MFVB}}$ defined via (11),

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The $\mathfrak{q}$-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 and Wand (2016) for the special case of $L = 0$ in the notation used there. However, as explained there, naïve updating of $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ has massive computational and storage costs when the number of groups is large. Result 1 asserts that we can instead use SolveTwoLevelSparseLeastSquares (Algorithm A.2) to obtain $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and relevant sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$.

B.3 Derivation of Result 2

Note that

$$\begin{aligned}
\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}) &\propto m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\, m_{(\boldsymbol{\beta},\boldsymbol{u})\,\rightarrow\,\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u}) \\
&= \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T)\end{bmatrix}\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}\right\} \\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}^T \boldsymbol{a} - \tfrac{1}{2}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}^T \boldsymbol{A}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{u}\end{bmatrix}\right\}
\end{aligned}$$

where $\boldsymbol{a}$ and $\boldsymbol{A}$ as given in Result 2 and 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

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

Result 2 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 (12).

B.4 Derivation of Algorithm 2

The two-level reduced exponential family form is

$$\begin{aligned}
\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}) \;&\propto\; \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i \\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T) \\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T) \end{bmatrix} \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}\right\} \\
&=\; \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{a} - \tfrac{1}{2}\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}^T \boldsymbol{A}\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}\right\}
\end{aligned}$$

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

$$\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}\boldsymbol{a} \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{A}^{-1}. \tag{36}$$

The first part of Algorithm 2 is such that the entries of $\boldsymbol{\eta}_{\mathfrak{q}(\boldsymbol{\beta},\,\boldsymbol{u})}$ are sequentially unpacked and stored in the vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_{4i}$, $1 \le i \le m$, corresponding to the $\boldsymbol{a}$ vector according to the partitioning in (27) and the matrices $\boldsymbol{\Omega}_3$ and $\boldsymbol{\Omega}_{7i}, \boldsymbol{\Omega}_{8i}$, $1 \le i \le m$, corresponding to the non-zero sub-blocks of $\boldsymbol{A}$ in (27).

Next, $\mathcal{S}_2$ stores the streamlined solution to (36) according to the SolveTwoLevelSparseMatrix algorithm (Algorithm A.1). The remainder of Algorithm 2 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 (36) and keeping in mind that (36) represents a two-level sparse matrix problem.

B.5 Derivation of Algorithm 3

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

$$
\begin{aligned}
\log\,\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^m \|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}-\boldsymbol{Z}_i\boldsymbol{u}_i\|^2+\text{const}\\
&=(1/\sigma^2)\begin{bmatrix}\boldsymbol{\beta}\\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T)\end{bmatrix}\end{bmatrix}^T\begin{bmatrix}\sum_{i=1}^m\boldsymbol{X}_i^T\boldsymbol{y}_i\\ -\frac{1}{2}\sum_{i=1}^m\boldsymbol{D}_p^T\text{vec}(\boldsymbol{X}_i^T\boldsymbol{X}_i)\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{Z}_i^T\boldsymbol{y}_i\\ -\frac{1}{2}\boldsymbol{D}_q^T\text{vec}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)\\ -\text{vec}(\boldsymbol{X}_i^T\boldsymbol{Z}_i)\end{bmatrix}\end{bmatrix}+\text{const.}
\end{aligned}
$$

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

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

where

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

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

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

The other factor to stochastic node message update is

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

where

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

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

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

Then note that

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

where, for example, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})} \equiv E_{\mathfrak{q}}(\boldsymbol{\beta})$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \equiv \text{Cov}_{\mathfrak{q}}(\boldsymbol{u}_i)$. Result 2 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 $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and sub-blocks of $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$. 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 TwoLevelNaturalToCommonParameters algorithm (Algorithm 2).

B.6 Derivation of Result 3

The derivation of Result 3 is very similar to that for Result 2.

B.7 Derivation of Algorithm 4

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

$$\begin{aligned}\log\,\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}) &= -\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}) - \frac{1}{2}\sum_{i=1}^m \boldsymbol{u}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_i + \text{const}\\ &= \begin{bmatrix}\boldsymbol{\beta}\\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i\\ \text{vech}(\boldsymbol{u}_i\boldsymbol{u}_i^T)\\ \text{vec}(\boldsymbol{\beta}\boldsymbol{u}_i^T)\end{bmatrix}\end{bmatrix}^T \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\\ -\tfrac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{0}_q\\ -\tfrac{1}{2}\boldsymbol{D}_q^T\text{vec}(\boldsymbol{\Sigma}^{-1})\\ \boldsymbol{0}_{pq}\end{bmatrix}\end{bmatrix} + \text{const.}\end{aligned}$$

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

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

where

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

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

$$\boldsymbol{M}_{\mathfrak{q}(\boldsymbol{\Sigma}^{-1})} \longleftarrow \{\omega_{12}+\tfrac{1}{2}(q+1)\}\{\mathrm{vec}^{-1}(\boldsymbol{\omega}_{13})\}^{-1}$$

where ω_{12} is the first entry of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma})\,\leftrightarrow\,\boldsymbol{\Sigma}}$ and $\boldsymbol{\omega}_{13}$ 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

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

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 $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)} \equiv \text{Cov}_{\mathfrak{q}}(\boldsymbol{u}_i)$. Result 3 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 2 to obtain $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i)}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i)}$, $1 \le i \le m$.

B.8 Derivation of Result 4

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

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

and

$$\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow (\boldsymbol{C}^T\boldsymbol{R}_{\text{MFVB}}^{-1}\boldsymbol{C} + \boldsymbol{D}_{\text{MFVB}})^{-1},$$

with $\boldsymbol{C}$, $\boldsymbol{D}_{\text{MFVB}}$ and $\boldsymbol{R}_{\text{MFVB}}$ as defined by

$$\boldsymbol{C} \equiv [\boldsymbol{X}\ \boldsymbol{Z}],\ \boldsymbol{D}_{\text{BLUP}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \underset{1\le i\le m}{\text{blockdiag}}\begin{bmatrix} \boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{n_i}\otimes\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}\end{bmatrix}\end{bmatrix}$$

and $\boldsymbol{R}_{\text{BLUP}} \equiv \sigma^2\boldsymbol{I}$ may be written as

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

where $\boldsymbol{B}$ and $\boldsymbol{b}$ have the sparse three-level forms given by (35) with

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

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

B.9 Derivation of Algorithm 5

Algorithm 5 is the three-level counterpart of Algorithm 1 and its derivation is analogous to that given for Algorithm 1 in Section B.2.

The first difference is that the $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u})}$ updates are expressible as three-level sparse matrix least squares problems and so the SolveThreeLevelSparseLeastSquares algorithm (Algorithm A.4) is used for streamlined updating of their relevant sub-blocks.

We still have $\mathfrak{q}(\sigma^2)$ optimally being an Inverse Chi-Squared density function but with shape parameter

$$\xi_{\mathfrak{q}(\sigma^2)} = \nu_{\sigma^2} + \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} o_{ij}$$

and rate parameter

$$\begin{aligned}\lambda_{\mathfrak{q}(\sigma^2)} =& \mu_{\mathfrak{q}(1/a_{\sigma^2})} + \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} E_{\mathfrak{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\}\\
=& \mu_{\mathfrak{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}_{\mathfrak{q}(\boldsymbol{\beta})} - \boldsymbol{Z}_{ij}^{\mathrm{L1}}\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})} - \boldsymbol{Z}_{ij}^{\mathrm{L2}}\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\|^2\\
&+\mathrm{tr}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})}) + \mathrm{tr}\{(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\} + \mathrm{tr}\{(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\}\\
&+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]\\
&+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]\\
&+2\,\mathrm{tr}\big[(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{Z}_{ij}^{\mathrm{L1}}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]\Big).\end{aligned}$$

The optimal $\mathfrak{q}(a_{\sigma^2})$ 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

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

where $\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} = \nu_{\boldsymbol{\Sigma}^{\mathrm{L1}}} + 2q_1 - 2 + m$ and

$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L1}})} = \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}}^{-1})} + \sum_{i=1}^{m}\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\mathrm{L1}})}\Big),$$

whilst

$$\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}}) \text{ is an Inverse-G-Wishart } \Big(G_{\mathrm{full}}, \xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}, \boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})}\Big) \text{ density function}$$

where $\xi_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} = \nu_{\boldsymbol{\Sigma}^{\mathrm{L2}}} + 2q_2 - 2 + \sum_{i=1}^m n_i$ and

$$\boldsymbol{\Lambda}_{\mathfrak{q}(\boldsymbol{\Sigma}^{\mathrm{L2}})} = \boldsymbol{M}_{\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}}^{-1})} + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big(\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}^T + \boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\mathrm{L2}})}\Big).$$

The optimal $\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L1}}})$ and $\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}^{\mathrm{L2}}})$ density functions have the same derivations and forms as $\mathfrak{q}(\boldsymbol{A}_{\boldsymbol{\Sigma}})$ in the two-level case.

Algorithm 5 is a streamlined iterative coordinate ascent for determination of Kullback-Leibler optimal values of each of the $\mathfrak{q}$-density parameters in the Bayesian three-level mixed model (21).

B.10 Derivation of Algorithm 6

Algorithm 6 is the three-level counterpart of Algorithm 2 and they each use the same logic. Therefore, the Algorithm 6 follows from arguments similar to those given in Section B.4.

B.11 Derivation of Result 5

Note that

$$
\begin{aligned}
\mathfrak{q}(\boldsymbol{\beta},\boldsymbol{u}) &\propto m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\, m_{(\boldsymbol{\beta},\boldsymbol{u})\,\to\,\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)}(\boldsymbol{\beta},\boldsymbol{u})\\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \mathrm{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\mathrm{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \mathrm{vech}\big(\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_i^{\mathrm{L1}})^T\big)\\ \mathrm{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_i^{\mathrm{L1}})^T\big)\end{bmatrix}\\ \underset{1\le i\le m}{\mathrm{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\mathrm{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\mathrm{L2}}\\ \mathrm{vech}\big(\boldsymbol{u}_{ij}^{\mathrm{L2}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\\ \mathrm{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\\ \mathrm{vec}\big(\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\end{bmatrix}\end{bmatrix}\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}\right\}\\
&= \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^T \boldsymbol{a} - \tfrac{1}{2}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^T \boldsymbol{A}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right\}
\end{aligned}
$$

where $\boldsymbol{a}$ and $\boldsymbol{A}$ are as given in Result 5. 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

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

and Result 5 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 $\boldsymbol{A}$.

B.12 Derivation of Algorithm 7

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

$$
\log\,\mathfrak{p}(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\|\boldsymbol{y}_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\beta}-\boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{u}_i^{\mathrm{L1}}-\boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{u}_{ij}^{\mathrm{L2}}\|^2+\text{const}
$$

$$
=(1/\sigma^2)\begin{bmatrix}\boldsymbol{\beta}\\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \text{vech}\big(\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_i^{\mathrm{L1}})^T\big)\\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_i^{\mathrm{L1}})^T\big)\end{bmatrix}\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\mathrm{L2}}\\ \text{vech}\big(\boldsymbol{u}_{ij}^{\mathrm{L2}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\mathrm{L1}})^T\big)\\ \text{vec}\big(\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\end{bmatrix}\end{bmatrix}\end{bmatrix}^T\boldsymbol{\nu}_1+\text{const.}
$$

where

$$
\boldsymbol{\nu}_1\equiv\begin{bmatrix}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{X}_{ij}^T\boldsymbol{y}_{ij}\\ \displaystyle-\tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij})\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\displaystyle\sum_{j=1}^{n_i}(\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{y}_{ij}\\ \displaystyle-\tfrac{1}{2}\sum_{j=1}^{n_i}\boldsymbol{D}_{q_1}^T\text{vec}\big((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L1}}\big)\\ \displaystyle-\sum_{j=1}^{n_i}\text{vec}\big(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\mathrm{L1}}\big)\end{bmatrix}\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix}(\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{y}_{ij}\\ -\tfrac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}((\boldsymbol{Z}_{ij}^{\mathrm{L2}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\\ -\text{vec}(\boldsymbol{X}_{ij}^T\boldsymbol{Z}_{ij}^{\mathrm{L2}})\\ -\text{vec}\big((\boldsymbol{Z}_{ij}^{\mathrm{L1}})^T\boldsymbol{Z}_{ij}^{\mathrm{L2}}\big)\end{bmatrix}\end{bmatrix}\end{bmatrix}.
$$

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

$$
\begin{aligned}
&m_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})\\
&\longleftarrow \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \mathrm{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\mathrm{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\mathrm{L1}}\\ \mathrm{vech}\big(\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_i^{\mathrm{L1}})^T\big)\\ \mathrm{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_i^{\mathrm{L1}})^T\big)\end{bmatrix}\\ \underset{1\le i\le m}{\mathrm{stack}}\left[\underset{1\le j\le n_i}{\mathrm{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\mathrm{L2}}\\ \mathrm{vech}\big(\boldsymbol{u}_{ij}^{\mathrm{L2}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\\ \mathrm{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\mathrm{L1}})^T\big)\\ \mathrm{vec}\big(\boldsymbol{u}_i^{\mathrm{L1}}(\boldsymbol{u}_{ij}^{\mathrm{L2}})^T\big)\end{bmatrix}\right]\end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})}\right\}
\end{aligned}
$$

where

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,(\boldsymbol{\beta},\boldsymbol{u})} \equiv \mu_{\mathfrak{q}(1/\sigma^2)}\,\boldsymbol{\nu}_1
$$

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

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

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

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

The other factor to stochastic node message update is

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

where

$$
\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\to\,\sigma^2} \equiv \begin{bmatrix} -\frac{1}{2}\displaystyle\sum_{i=1}^m\sum_{j=1}^{n_i} o_{ij}\\ -\frac{1}{2}\displaystyle\sum_{i=1}^m E_{\mathfrak{q}}\{\|\boldsymbol{y}_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\beta}-\boldsymbol{Z}_{ij}^{\mathrm{L1}}\boldsymbol{u}_i^{\mathrm{L1}}-\boldsymbol{Z}_{ij}^{\mathrm{L2}}\boldsymbol{u}_{ij}^{\mathrm{L2}}\|^2\}\end{bmatrix}
$$

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

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

Observing that

$$
\begin{aligned}
&E_{\mathfrak{q}}\{\|\boldsymbol{y}_{ij}-\boldsymbol{X}_{ij}\boldsymbol{\beta}-\boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{u}_{ij}^{\text{L2}}\|^2\}\\
&\quad=\|\boldsymbol{y}_{ij}-\boldsymbol{X}_{ij}\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})}-\boldsymbol{Z}_{ij}^{\text{L1}}\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}-\boldsymbol{Z}_{ij}^{\text{L2}}\,\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\|^2+\text{tr}(\boldsymbol{X}_{ij}^T\boldsymbol{X}_{ij}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{\beta})})\\
&\qquad+\text{tr}\{(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{Z}_{ij}^{\text{L1}}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\}+\text{tr}\{(\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{Z}_{ij}^{\text{L2}}\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\}\\
&\qquad+2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{L1}})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})^T\}\big]\\
&\qquad+2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{X}_{ij}E_{\mathfrak{q}}\{(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{\beta})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\big]\\
&\qquad+2\,\text{tr}\big[(\boldsymbol{Z}_{ij}^{\text{L2}})^T\boldsymbol{Z}_{ij}^{\text{L1}}E_{\mathfrak{q}}\{(\boldsymbol{u}_i^{\text{L1}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})})(\boldsymbol{u}_{ij}^{\text{L2}}-\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})})^T\}\big]
\end{aligned}
$$

Result 5 shows how the sub-blocks of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{y}|\,\boldsymbol{\beta},\boldsymbol{u},\sigma^2)\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ 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

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

in the call to ThreeLevelNaturalToCommonParameters algorithm (Algorithm 6).

B.13 Derivation of Result 6

The derivation of Result 6 is very similar to that for Result 5.

B.14 Derivation of Algorithm 8

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

$$
\begin{aligned}
\log\,\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})&=\tfrac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})-\tfrac{1}{2}\sum_{i=1}^m(\boldsymbol{u}_i^{\text{L1}})^T(\boldsymbol{\Sigma}^{\text{L1}})^{-1}\boldsymbol{u}_i^{\text{L1}}\\
&\quad-\tfrac{1}{2}\sum_{i=1}^m\sum_{j=1}^{n_i}(\boldsymbol{u}_{ij}^{\text{L2}})^T(\boldsymbol{\Sigma}^{\text{L2}})^{-1}\boldsymbol{u}_{ij}^{\text{L2}}+\text{const}\\
&=\begin{bmatrix}\boldsymbol{\beta}\\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_i^{\text{L1}}\\ \text{vech}\big(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T\big)\\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T\big)\end{bmatrix}\\ \underset{1\le i\le m}{\text{stack}}\left[\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix}\boldsymbol{u}_{ij}^{\text{L2}}\\ \text{vech}\big(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big)\\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big)\\ \text{vec}\big(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big)\end{bmatrix}\right]\end{bmatrix}^T\boldsymbol{\nu}_2+\text{const}
\end{aligned}
$$

where

$$\boldsymbol{\nu}_2 \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ -\frac{1}{2}\boldsymbol{D}_p^T\text{vec}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \mathbf{0}_{q_1} \\ -\frac{1}{2}\boldsymbol{D}_{q_1}^T\text{vec}\big((\boldsymbol{\Sigma}^{\text{L1}})^{-1}\big) \\ \mathbf{0}_{pq_1} \end{bmatrix} \\ \underset{1\le i\le m}{\text{stack}}\left[\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix} \mathbf{0}_{q_2} \\ -\frac{1}{2}\boldsymbol{D}_{q_2}^T\text{vec}\big((\boldsymbol{\Sigma}^{\text{L2}})^{-1}\big) \\ \mathbf{0}_{pq_2} \\ \mathbf{0}_{q_1q_2} \end{bmatrix}\right] \end{bmatrix}.$$

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

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}(\boldsymbol{\beta},\boldsymbol{u})$$
$$\longleftarrow \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \text{vech}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \\ \underset{1\le i\le m}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_i^{\text{L1}} \\ \text{vech}\big(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T\big) \\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_i^{\text{L1}})^T\big) \end{bmatrix} \\ \underset{1\le i\le m}{\text{stack}}\left[\underset{1\le j\le n_i}{\text{stack}}\begin{bmatrix} \boldsymbol{u}_{ij}^{\text{L2}} \\ \text{vech}\big(\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big) \\ \text{vec}\big(\boldsymbol{\beta}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big) \\ \text{vec}\big(\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\big) \end{bmatrix}\right] \end{bmatrix}^T \boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,(\boldsymbol{\beta},\boldsymbol{u})} \equiv \boldsymbol{\nu}_2.$$

Here $\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})}$ denotes expectation of $(\boldsymbol{\Sigma}^{\text{L1}})^{-1}$ with respect to the normalization of

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

which is an Inverse G-Wishart density function with natural parameter vector

$$\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}$$

and, according to Table S.1 in the online supplement of Wand (2017), leads to

$$\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L1}})^{-1})} \longleftarrow \{\omega_{36} + \tfrac{1}{2}(q_1+1)\}\{\text{vec}^{-1}(\boldsymbol{\omega}_{37})\}^{-1}$$

where ω_{36} is the first entry of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}$ and $\boldsymbol{\omega}_{37}$ is the vector containing the remaining entries of $\eta_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\leftrightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}$. The treatment of $\boldsymbol{M}_{\mathfrak{q}((\boldsymbol{\Sigma}^{\text{L2}})^{-1})}$ is analogous.

The message from $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})$ to $\boldsymbol{\Sigma}^{\text{L1}}$ is

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}(\boldsymbol{\Sigma}^{\text{L1}})\longleftarrow\exp\left\{\left[\begin{array}{c}\log|\boldsymbol{\Sigma}^{\text{L1}}|\\[1ex]\text{vech}\big((\boldsymbol{\Sigma}^{\text{L1}})^{-1}\big)\end{array}\right]^T\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L1}}}\equiv\left[\begin{array}{c}-\frac{1}{2}\,m\\[2ex]-\frac{1}{2}\sum_{i=1}^m\boldsymbol{D}_{q_1}^T\text{vec}[E_{\mathfrak{q}}\{\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T\}]\end{array}\right]$$

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

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

Similarly, the message from $\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})$ to $\boldsymbol{\Sigma}^{\text{L2}}$ is

$$m_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L2}}}(\boldsymbol{\Sigma}^{\text{L2}})\longleftarrow\exp\left\{\left[\begin{array}{c}\log|\boldsymbol{\Sigma}^{\text{L2}}|\\[1ex]\text{vech}\big((\boldsymbol{\Sigma}^{\text{L2}})^{-1}\big)\end{array}\right]^T\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L2}}}\right\}$$

where

$$\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\rightarrow\,\boldsymbol{\Sigma}^{\text{L2}}}\equiv\left[\begin{array}{c}-\frac{1}{2}\sum_{i=1}^m n_i\\[2ex]-\frac{1}{2}\sum_{i=1}^m\sum_{j=1}^{n_i}\boldsymbol{D}_{q_2}^T\text{vec}[E_{\mathfrak{q}}\{\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\}]\end{array}\right].$$

Now note that

$$E_{\mathfrak{q}}\{\boldsymbol{u}_i^{\text{L1}}(\boldsymbol{u}_i^{\text{L1}})^T\}=\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\quad\text{and}\quad E_{\mathfrak{q}}\{\boldsymbol{u}_{ij}^{\text{L2}}(\boldsymbol{u}_{ij}^{\text{L2}})^T\}=\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}^T+\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$$

where, similar to before, $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\equiv E_{\mathfrak{q}}(\boldsymbol{u}_i^{\text{L1}})$, $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}\equiv\text{Cov}_{\mathfrak{q}}(\boldsymbol{u}_i^{\text{L1}})$ and $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ is defined similarly. Result 6 links sub-blocks of $\boldsymbol{\eta}_{\mathfrak{p}(\boldsymbol{\beta},\boldsymbol{u}|\,\boldsymbol{\Sigma}^{\text{L1}},\boldsymbol{\Sigma}^{\text{L2}})\,\leftrightarrow\,(\boldsymbol{\beta},\boldsymbol{u})}$ 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 6 to obtain $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_i^{\text{L1}})}$, $1\le i\le m$, as well as $\boldsymbol{\mu}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$ and $\boldsymbol{\Sigma}_{\mathfrak{q}(\boldsymbol{u}_{ij}^{\text{L2}})}$, $1\le i\le m$, $1\le j\le n_i$.

Algorithm 8 is a proceduralization of each of these results.

References

A. Atay-Kayis and H. Massam. A Monte Carlo method for computing marginal likelihood in nondecomposable gaussian graphical models. *Biometrika*, 92:317–335, 2005.

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

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

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

J.B. Duck-Mayr. *RcppDist: Rcpp integration of additional probability distributions*, 2018. URL https://CRAN.R-project.org. R package version 0.1.1.

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

D. Eddelbeuttel, R. Francois, J.J. Allaire, K. Ushey, Q. Kou, N. Russell, D. Bates, and J. Chambers. *Rcpp: Seamless R and C++ integration*, 2020a. URL https://www.rcpp.org. R package version 1.0.5.

D. Eddelbeuttel, R. Francois, D. Bates, and B. Ni. *RcppArmadillo: Rcpp integration for the Armadillo templated linear algebra library*, 2020b. URL https://CRAN.R-project.org. R package version 0.9.900.2.0.

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

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

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

G.H. Golub and C.F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1989.

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

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

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

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

L. Maestrini and M.P. Wand. The Inverse G-Wishart distribution and variational message passing. 2020. URL https://arxiv.org/abs/2005.09876.

M. Menictas and M.P. Wand. Variational inference for marginal longitudinal semiparametric regression. *Stat*, 2:61–71, 2013.

M. Menictas, T.H. Nolan, D.G. Simpson, and M.P. Wand. Streamlined variational inference for higher level group-specific curve models. *Statistical Modelling*, 2020. in press.

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

T.H. Nolan and M.P. Wand. Streamlined solutions to multilevel sparse matrix problems. *ANZIAM Journal*, 62:18–41, 2020.

J.T. Ormerod and M.P. Wand. Explaining variational approximations. *The American Statistician*, 64:140–153, 2010.

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

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

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

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

R Core Team. *R: A language and environment for statistical computing.* Vienna, Austria, 2020. URL https://www.R-project.org. R Foundation for Statistical Computing.

Stan Development Team. *RStan: RStan: the R interface to Stan*, 2020. URL https://mc-stan.org. R package version 2.21.2.

M.P. Wand. Fast approximate inference for arbitrarily large semiparametric regression models via message passing (with discussion). *Journal of the American Statistical Association*, 112:137–168, 2017.

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

B. Wang and D.M. Titterington. Inadequacy of interval estimates corresponding to variational Bayes approximations. In Z. Ghahramani and R. Cowell, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pages 373–380, 2005.