

Streamlined Variational Inference for Higher Level Group-Specific Curve Models

BY M. MENICTAS¹, T.H. NOLAN¹, D.G. SIMPSON² AND M.P. WAND¹

University of Technology Sydney¹ and University of Illinois²

11th March, 2019

Abstract

A two-level group-specific curve model is such that the mean response of each member of a group is a separate smooth function of a predictor of interest. The three-level extension is such that one grouping variable is nested within another one, and higher level extensions are analogous. Streamlined variational inference for higher level group-specific curve models is a challenging problem. We confront it by systematically working through two-level and then three-level cases and making use of the higher level sparse matrix infrastructure laid down in Nolan & Wand (2018). A motivation is analysis of data from ultrasound technology for which three-level group-specific curve models are appropriate. Whilst extension to the number of levels exceeding three is not covered explicitly, the pattern established by our systematic approach sheds light on what is required for even higher level group-specific curve models.

Keywords: longitudinal data analysis, multilevel models, panel data, mean field variational Bayes.

1 Introduction

We provide explicit algorithms for fitting and approximate Bayesian inference for multi-level models involving, potentially, thousands of noisy curves. The algorithms include covariance parameter estimation and allow for pointwise credible intervals around the fitted curves. Contrast function fitting and inference is also supported by our approach. Both two-level and three-level situations are covered, and a template for even higher level situations is laid down.

Models and methodology for statistical analyses of grouped data for which the basic unit is a noisy curve continues to be an important area of research. A driving force is rapid technological change which is resulting in the generation of curve-type data at fine resolution levels. Examples of such technology include accelerometers (e.g. Goldsmith *et al.*, 2015) personal digital assistants (e.g. Trail *et al.*, 2014) and quantitative ultrasound (e.g. Wirtzfeld *et al.*, 2015). In some applications curve-type data have higher levels of grouping, with groups at one level nested inside other groups. Our focus here is streamlined variational inference for such circumstances.

Some motivating data is shown in Figure 1 from an experiment involving quantitative ultrasound technology. Each curve corresponds to a logarithmically transformed backscatter coefficient over a fine grid of frequency values for tumors in laboratory mice, with exactly one tumor per mouse. The backscatter/frequency curves are grouped according to one of 5 slices of the same tumor, corresponding to probe locations. The slices are grouped according to being from one of 10 tumors. We refer to such data as three-level data with frequency measurements at level 1, slices being the level 2 groups and tumors constituting the level 3 groups. The gist of this article is efficient and flexible variational fitting and inference for such data, that scales well to much larger multilevel data sets. Indeed, our algorithms are linear in the number of groups at both level 2 and level 3. Simulation study

results given later in this article show that curve-type data with thousands of groups can be analyzed quickly using our new methodology. Depending on sample sizes and implementation language, fitting times range from a few seconds to a few minutes. In contrast, naïve implementations become infeasible when the number of groups are in the several hundreds due to storage and computational demands.

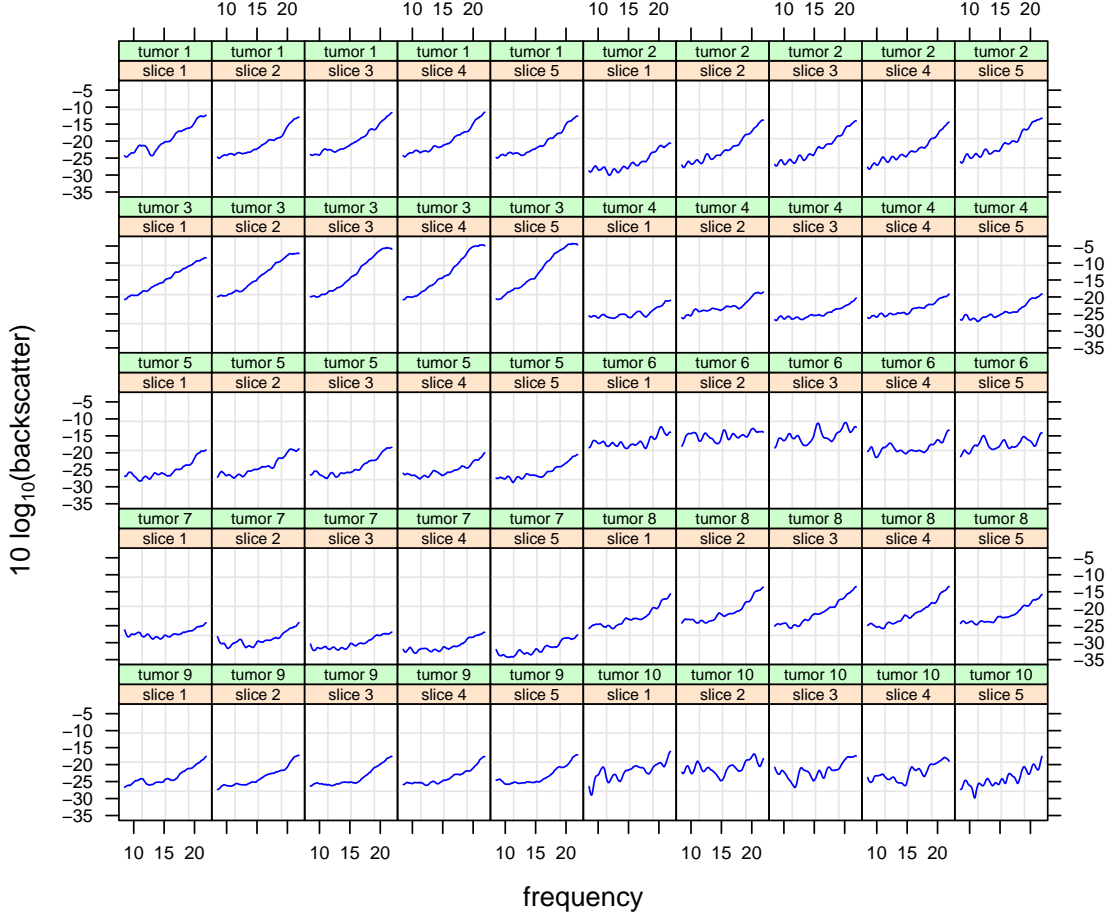


Figure 1: Illustrative three-level curve-type data. The response variable is $10 \log_{10}(\text{backscatter})$ according to ultrasound technology. Level 1 corresponds to different ultrasound frequencies and matches the horizontal axes in each panel. Level 2 corresponds to different slices of a tumor due to differing probe locations. Level 3 corresponds to different tumors with one tumor for each of 10 laboratory mice.

We work with a variant of group-specific curve models that at least go back to Donnelly, Laird & Ware (1995). Other contributions of this type include Brumback & Rice (1998), Verbyla *et al.* (1999), Wang (1998) and Zhang *et al.* (1998). The specific formulation that we use is that given by Durban *et al.* (2005) which involves an embedding within the class of linear mixed models (e.g. Robinson, 1991) with low-rank smoothing splines used for flexible function modelling and fitting.

Even though approximate Bayesian variational inference is our overarching goal, we also provide an important parallelism involving classical frequentist inference. Contemporary mixed model software such as `nlme()` (Pinheiro *et al.*, 2018) and `lme4()` (Bates *et al.*, 2015) in the R language provide streamlined algorithms for obtaining the best linear unbiased predictions of fixed and random effects in multilevel mixed models with details given in, for example, Pinheiro & Bates (2000). However, the sub-blocks of the covariance matrices required for construction of pointwise confidence interval bands around the esti-

mated curves are *not* provided by such software. In the variational Bayesian analog, these sub-blocks are required for covariance parameter fitting and inference which, in turn, are needed for curve estimation. A significant contribution of this article is streamlined computation for both the best linear unbiased predictors and its corresponding covariance computation. Similar mathematical results lead to the mean field variational Bayesian inference equivalent. We present explicit ready-to-code algorithms for both two-level and three-level group-specific curve models. Extensions to higher level models could be derived using the blueprint that we establish here. Nevertheless, the algebraic overhead is increasingly burdensome with each increment in the number of levels. It is prudent to treat each multilevel case separately and here we already require several pages to cover two-level and three-level group-specific curve models. To our knowledge, this is the first article to provide streamlined algorithms for fitting three-level group-specific curve models.

Another important aspect of our group-specific curve fitting algorithms is the fact that they make use of the `SOLVETWOLEVELSPARSELEASTSQUARES` and `SOLVETHREELEVELSPARSELEASTSQUARES` algorithms developed for ordinary linear mixed models in Nolan *et al.* (2018). This realization means that the algorithms listed in Sections 2 and 3 are more concise and code-efficient: there is no need to repeat the implementation of these two fundamental algorithms for stable QR-based solving of higher level sparse linear systems. Sections S.11–S.12 of the web-supplement provide details on the `SOLVETHREELEVELSPARSELEASTSQUARES` and `SOLVETHREELEVELSPARSELEASTSQUARES` algorithms.

Section 2 deals with the two-level case and the three-level case is covered in Section 3. In Section 4 we provide some assessments concerning the accuracy and speed of the new variational inference algorithms.

2 Two-Level Models

The simplest version of group-specific curve models involves the pairs (x_{ij}, y_{ij}) where x_{ij} is the j th value of the predictor variable within the i th group and y_{ij} is the corresponding value of the response variable. We let m denote the number of groups and n_i denote the number of predictor/response pairs within the i th group. The Gaussian response two-level group specific curve model is

$$y_{ij} = f(x_{ij}) + g_i(x_{ij}) + \varepsilon_i, \quad \varepsilon_{ij} \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i, \quad (1)$$

where the smooth function f is the global regression mean function and the smooth functions g_i , $1 \leq i \leq m$, allow for flexible group-specific deviations from f . As in Durban *et al.* (2005), we use mixed model-based penalized basis functions to model f and the g_i . Specifically,

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k} z_{\text{gbl},k}(x), \quad u_{\text{gbl},k} \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{gbl}}^2), \text{ and}$$

$$g_i(x) = u_{\text{lin},i0} + u_{\text{lin},i1} x + \sum_{k=1}^{K_{\text{grp}}} u_{\text{grp},ik} z_{\text{grp},k}(x), \quad \begin{bmatrix} u_{\text{lin},i0} \\ u_{\text{lin},i1} \end{bmatrix} \stackrel{\text{ind.}}{\sim} N(\mathbf{0}, \Sigma), \quad u_{\text{grp},ik} \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{grp}}^2),$$

where $\{z_{\text{gbl},k}(\cdot) : 1 \leq k \leq K_{\text{gbl}}\}$ and $\{z_{\text{grp},k}(\cdot) : 1 \leq k \leq K_{\text{grp}}\}$ are suitable sets of basis functions. Splines and wavelet families are the most common choices for the $z_{\text{gbl},k}(\cdot)$ and $z_{\text{grp},k}(\cdot)$. In our illustrations and simulation studies we use the canonical cubic O’Sullivan spline basis as described in Section 4 of Wand & Ormerod (2008), which corresponds to a low-rank version of classical smoothing splines (e.g. Wahba, 1990). The variance parameters σ_{gbl}^2 and σ_{grp}^2 control the effective degrees of freedom used for the global mean and

group-specific deviation functions respectively. Lastly, Σ is a 2×2 unstructured covariance matrix for the coefficients of the group-specific linear deviations.

We also use the notation:

$$\mathbf{x}_i \equiv \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{bmatrix} \quad \text{and} \quad \mathbf{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}$$

for the vectors of predictors and responses corresponding to the i th group. Notation such as $z_{\text{gbl},1}(\mathbf{x}_i)$ denotes the $n_i \times 1$ vector containing $z_{\text{gbl},1}(x_{ij})$ values, $1 \leq j \leq n_i$.

2.1 Best Linear Unbiased Prediction

Model (1) is expressible as a Gaussian response linear mixed model as follows:

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2 \mathbf{I}), \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}), \quad (2)$$

where

$$\mathbf{X} \equiv \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} \quad \text{with} \quad \mathbf{X}_i \equiv [\mathbf{1} \ \mathbf{x}_i] \quad \text{and} \quad \boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

are the fixed effects design matrix and coefficients, corresponding to the linear component of f . The random effects design matrix \mathbf{Z} and corresponding random effects vector \mathbf{u} are partitioned according to

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{\text{gbl}} & \text{blockdiag}([\mathbf{X}_i \ \mathbf{Z}_{\text{grp},i}]) \\ & 1 \leq i \leq m \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_{\text{gbl}} \\ \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix}_{1 \leq i \leq m} \end{bmatrix} \quad (3)$$

where $\mathbf{u}_{\text{gbl}} = [u_{\text{gbl},1} \ \dots \ u_{\text{gbl},K_{\text{gbl}}}]^T$ are the coefficients corresponding to the non-linear component of f , $\mathbf{u}_{\text{lin},i} = [u_{\text{lin},i0} \ u_{\text{lin},i1}]^T$ are the coefficients corresponding to the linear component of g_i and $\mathbf{u}_{\text{grp},i} = [u_{\text{grp},i1} \ \dots \ u_{\text{grp},iK_{\text{grp}}}]^T$ are the coefficients corresponding to the non-linear component of g_i , $1 \leq i \leq m$. In (3), $\mathbf{Z}_{\text{gbl}} \equiv \text{stack}_{1 \leq i \leq m}(\mathbf{Z}_{\text{gbl},i})$ and the matrices $\mathbf{Z}_{\text{gbl},i}$ and $\mathbf{Z}_{\text{grp},i}$, $1 \leq i \leq m$, contain, respectively, spline basis functions for the global mean function f and the i th group deviation functions g_i . Specifically,

$$\mathbf{Z}_{\text{gbl},i} \equiv [z_{\text{gbl},1}(\mathbf{x}_i) \ \dots \ z_{\text{gbl},K_{\text{gbl}}}(\mathbf{x}_i)] \quad \text{and} \quad \mathbf{Z}_{\text{grp},i} = [z_{\text{grp},1}(\mathbf{x}_i) \ \dots \ z_{\text{grp},K_{\text{grp}}}(\mathbf{x}_i)]$$

for $1 \leq i \leq m$. The corresponding fixed and random effects vectors are

$$\mathbf{u}_{\text{gbl}} \sim N(\mathbf{0}, \sigma_{\text{gbl}}^2 \mathbf{I}_{K_{\text{gbl}}}) \quad \text{and} \quad \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \stackrel{\text{ind.}}{\sim} N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp}}^2 \mathbf{I}_{K_{\text{grp}}} \end{bmatrix}\right), \quad 1 \leq i \leq m.$$

Hence, the full random effects covariance matrix is

$$\mathbf{G} = \text{Cov}(\mathbf{u}) = \begin{bmatrix} \sigma_{\text{gbl}}^2 \mathbf{I}_{K_{\text{gbl}}} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_m \otimes \begin{bmatrix} \Sigma & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp}}^2 \mathbf{I}_{K_{\text{grp}}} \end{bmatrix} \end{bmatrix}. \quad (4)$$

Next define the matrices

$$\mathbf{C} \equiv [\mathbf{X} \ \mathbf{Z}], \quad \mathbf{D}_{\text{BLUP}} \equiv \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{G}^{-1} \end{bmatrix} \quad \text{and} \quad \mathbf{R}_{\text{BLUP}} \equiv \sigma_\varepsilon^2 \mathbf{I}. \quad (5)$$

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

$$\begin{aligned} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} \end{bmatrix} &= (\mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{C} + \mathbf{D}_{\text{BLUP}})^{-1} \mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{y} \\ \text{and } \text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}} - \mathbf{u} \end{bmatrix} \right) &= (\mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{C} + \mathbf{D}_{\text{BLUP}})^{-1}. \end{aligned} \quad (6)$$

This covariance matrix grows quadratically in m , so its storage becomes infeasible for large numbers of groups. However, only the following sub-blocks are required for adding pointwise confidence intervals to curve estimates:

$$\begin{aligned} \text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right) &= \text{top left-hand } (2 + K_{\text{gbl}}) \times (2 + K_{\text{gbl}}) \\ &\quad \text{sub-block of } (\mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{C} + \mathbf{D}_{\text{BLUP}})^{-1}, \\ \text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) &= \text{subsequent } (2 + K_{\text{grp}}) \times (2 + K_{\text{grp}}) \text{ diagonal} \\ &\quad \text{sub-blocks of } (\mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{C} + \mathbf{D}_{\text{BLUP}})^{-1} \\ &\quad \text{below Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right), 1 \leq i \leq m, \text{ and} \\ E \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix}^T \right\} &= \text{subsequent } (2 + K_{\text{gbl}}) \times (2 + K_{\text{grp}}) \text{ sub-blocks} \\ &\quad \text{of } (\mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{C} + \mathbf{D}_{\text{BLUP}})^{-1} \text{ to the right of} \\ &\quad \text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right), 1 \leq i \leq m. \end{aligned} \quad (7)$$

As in Nolan, Menictas & Wand (2019), we define the generic two-level sparse matrix to be determination of the vector \mathbf{x} which minimizes the least squares criterion

$$\|\mathbf{b} - \mathbf{B}\mathbf{x}\|^2 \quad \text{where } \|\mathbf{v}\|^2 \equiv \mathbf{v}^T \mathbf{v} \text{ for any column vector } \mathbf{v}, \quad (8)$$

with \mathbf{B} having the two-level sparse form

$$\mathbf{B} \equiv \begin{bmatrix} \mathbf{B}_1 & \dot{\mathbf{B}}_1 & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{B}_2 & \mathbf{O} & \dot{\mathbf{B}}_2 & \cdots & \mathbf{O} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_m & \mathbf{O} & \mathbf{O} & \cdots & \dot{\mathbf{B}}_m \end{bmatrix} \quad \text{and } \mathbf{b} \text{ partitioned according to } \mathbf{b} \equiv \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}. \quad (9)$$

In (9), for any $1 \leq i \leq m$, the matrices \mathbf{B}_i , $\dot{\mathbf{B}}_i$ and \mathbf{b}_i each have the same number of rows. The numbers of columns in \mathbf{B}_i and $\dot{\mathbf{B}}_i$ are arbitrary whereas the \mathbf{b}_i are column vectors. In addition to solving for \mathbf{x} , the sub-blocks of $(\mathbf{B}^T \mathbf{B})^{-1}$ corresponding to the non-sparse regions of $\mathbf{B}^T \mathbf{B}$ are included in our definition of a two-level sparse matrix least squares problem. Algorithm 2 of Nolan *et al.* (2018) provides a stable and efficient solution to this problem and labels it the SOLVETWOLEVELSPARSELEASTSQUARES algorithm. Section S.11 of the web-supplement contains details regarding this algorithm. In Nolan *et al.* (2018) we used SOLVETWOLEVELSPARSELEASTSQUARES for fitting two-level linear mixed models. However, precisely the same algorithm can be used for fitting two-level group-specific curve models because of:

Result 1. Computation of $[\hat{\beta}^T \hat{\mathbf{u}}^T]^T$ and each of the sub-blocks of $\text{Cov}([\hat{\beta}^T (\hat{\mathbf{u}} - \mathbf{u})^T]^T)$ listed in (7) are expressible as solutions to the two-level sparse matrix least squares problem:

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

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

$$\mathbf{b}_i \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{y}_i \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_i \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_i & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{gbl},i} \\ \mathbf{O} & m^{-1/2} \sigma_{\text{gbl}}^{-1} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \text{and} \quad \dot{\mathbf{B}}_i \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_i & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{grp},i} \\ \mathbf{O} & \mathbf{O} \\ \Sigma^{-1/2} & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp}}^{-1} \mathbf{I}_{K_{\text{grp}}} \end{bmatrix}$$

with each of these matrices having $\tilde{n}_i = n_i + K_{\text{gbl}} + 2 + K_{\text{grp}}$ rows and with \mathbf{B}_i having $p = 2 + K_{\text{gbl}}$ columns and $\dot{\mathbf{B}}_i$ having $q = 2 + K_{\text{grp}}$ columns. The solutions are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} \end{bmatrix} = \mathbf{x}_1, \quad \text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right) = \mathbf{A}^{11}$$

and

$$\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} \end{bmatrix} = \mathbf{x}_{2,i}, \quad E \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix}^T \right\} = \mathbf{A}^{12,i},$$

$$\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) = \mathbf{A}^{22,i}, \quad 1 \leq i \leq m.$$

A derivation of Result 1 is given in Section S.1 of the web-supplement. Algorithm 1 encapsulates streamlined best linear unbiased prediction computation together with coefficient covariance matrix sub-blocks of interest.

2.2 Mean Field Variational Bayes

We now consider the following Bayesian extension of (2) and (4):

$$\mathbf{y} | \beta, \mathbf{u}, \sigma_\varepsilon^2 \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2 \mathbf{I}), \quad \mathbf{u} | \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \Sigma \sim N(\mathbf{0}, \mathbf{G}), \quad \mathbf{G} \text{ as defined in (4)},$$

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

$$\sigma_{\text{gbl}}^2 | a_{\text{gbl}} \sim \text{Inverse-}\chi^2(\nu_{\text{gbl}}, 1/a_{\text{gbl}}), \quad a_{\text{gbl}} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\text{gbl}} s_{\text{gbl}}^2)),$$

$$\sigma_{\text{grp}}^2 | a_{\text{grp}} \sim \text{Inverse-}\chi^2(\nu_{\text{grp}}, 1/a_{\text{grp}}), \quad a_{\text{grp}} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\text{grp}} s_{\text{grp}}^2)),$$

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

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

(10)

Here the 2×1 vector $\boldsymbol{\mu}_\beta$ and 2×2 symmetric positive definite matrix Σ_β are hyperparameters corresponding to the prior distribution on β and

$$\nu_\varepsilon, s_\varepsilon, \nu_{\text{gbl}}, s_{\text{gbl}}, \nu_{\text{grp}}, s_{\text{grp}}, \nu_\Sigma, s_{\Sigma,1}, s_{\Sigma,2} > 0$$

are hyperparameters for the variance and covariance matrix parameters. Details on the Inverse G-Wishart distribution, and the Inverse- χ^2 special case, are given in Section S.3 of

Algorithm 1 Streamlined algorithm for obtaining best linear unbiased predictions and corresponding covariance matrix components for the two-level group specific curves model.

Inputs: $\mathbf{y}_i(n_i \times 1)$, $\mathbf{X}_i(n_i \times 2)$, $\mathbf{Z}_{\text{gbl},i}(n_i \times K_{\text{gbl}})$, $\mathbf{Z}_{\text{grp},i}(n_i \times K_{\text{grp}})$, $1 \leq i \leq m$; $\sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2 > 0$, $\Sigma(q \times q)$, symmetric and positive definite.

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

$$\mathbf{b}_i \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{y}_i \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_i \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_i & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{gbl},i} \\ \mathbf{O} & m^{-1/2} \sigma_{\text{gbl}}^{-1} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{\mathbf{B}}_i \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_i & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{grp},i} \\ \mathbf{O} & \mathbf{O} \\ \Sigma^{-1/2} & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp}}^{-1} \mathbf{I}_{K_{\text{grp}}} \end{bmatrix}$$

$\mathcal{S}_1 \leftarrow \text{SOLVETWOLEVELSPARSELEASTSQUARES}(\{(\mathbf{b}_i, \mathbf{B}_i, \dot{\mathbf{B}}_i) : 1 \leq i \leq m\})$

$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} \end{bmatrix} \leftarrow \mathbf{x}_1 \text{ component of } \mathcal{S}_1$; $\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right) \leftarrow \mathbf{A}^{11} \text{ component of } \mathcal{S}_1$

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

$$\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} \end{bmatrix} \leftarrow \mathbf{x}_{2,i} \text{ component of } \mathcal{S}_1$$

$$\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) \leftarrow \mathbf{A}^{22,i} \text{ component of } \mathcal{S}_1$$

$$E \left\{ \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix}^T \right\} \leftarrow \mathbf{A}^{12,i} \text{ component of } \mathcal{S}_1$$

Output:

$$\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} \end{bmatrix}, \text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right), \left\{ \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} \end{bmatrix}, \text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix} \right), \right. \right.$$

$$\left. E \left\{ \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix}^T \right\} : 1 \leq i \leq m \right\}$$

the web-supplement. The auxiliary variable a_ε is defined so that σ_ε has a Half- t distribution with degrees of freedom parameter ν_ε and scale parameter s_ε , with larger values of s_ε corresponding to greater noninformativity. Analogous comments apply to the other standard deviation parameters. Setting $\nu_\Sigma = 2$ leads to the correlation parameter in Σ having a Uniform distribution on $(-1, 1)$ (Huang & Wand, 2013).

Throughout this article we use \mathbf{p} generically to denote a density function corresponding to random quantities in Bayesian models such as (10). For example, $\mathbf{p}(\boldsymbol{\beta})$ denotes the prior density function of $\boldsymbol{\beta}$ and $\mathbf{p}(\mathbf{u} | \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \Sigma)$ denotes the density function of \mathbf{u} conditional on $(\sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \Sigma)$. Now consider the following mean field restriction on the joint posterior density function of all parameters in (10):

$$\mathbf{p}(\boldsymbol{\beta}, \mathbf{u}, a_\varepsilon, a_{\text{gbl}}, a_{\text{grp}}, \mathbf{A}_\Sigma, \sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \Sigma | \mathbf{y}) \approx \mathbf{q}(\boldsymbol{\beta}, \mathbf{u}, a_\varepsilon, a_{\text{gbl}}, a_{\text{grp}}, \mathbf{A}_\Sigma) \mathbf{q}(\sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \Sigma). \quad (11)$$

Here, generically, each q denotes an approximate posterior density function of the random vector indicated by its argument according to the mean field restriction (11). Then application of the minimum Kullback-Leibler divergence equations (e.g. equation (10.9) of Bishop, 2006) leads to the optimal q -density functions for the parameters of interest being as follows:

$$\begin{aligned} q^*(\beta, u) &\text{ has a } N(\mu_{q(\beta, u)}, \Sigma_{q(\beta, u)}) \text{ distribution,} \\ q^*(\sigma_\varepsilon^2) &\text{ has an Inverse-}\chi^2(\xi_{q(\sigma_\varepsilon^2)}, \lambda_{q(\sigma_\varepsilon^2)}) \text{ distribution,} \\ q^*(\sigma_{\text{gbl}}^2) &\text{ has an Inverse-}\chi^2(\xi_{q(\sigma_{\text{gbl}}^2)}, \lambda_{q(\sigma_{\text{gbl}}^2)}) \text{ distribution,} \\ q^*(\sigma_{\text{grp}}^2) &\text{ has an Inverse-}\chi^2(\xi_{q(\sigma_{\text{grp}}^2)}, \lambda_{q(\sigma_{\text{grp}}^2)}) \text{ distribution} \end{aligned}$$

$$\text{and } q^*(\Sigma) \text{ has an Inverse-G-Wishart}(G_{\text{full}}, \xi_{q(\Sigma)}, \Lambda_{q(\Sigma)}) \text{ distribution.}$$

The optimal q -density parameters are determined via an iterative coordinate ascent algorithm, with details given in Section S.5 of this article's web-supplement. The stopping criterion is based on the variational lower bound on the marginal likelihood (e.g. Bishop, 2006; Section 10.2.2) and denoted $p(y; q)$. Its logarithmic form and derivation are given in Section S.6 of the web-supplement.

Note that updates for $\mu_{q(\beta, u)}$ and $\Sigma_{q(\beta, u)}$ may be written

$$\mu_{q(\beta, u)} \leftarrow (C^T R_{\text{MFVB}}^{-1} C + D_{\text{MFVB}})^{-1} (C^T R_{\text{MFVB}}^{-1} y + o_{\text{MFVB}}) \quad \text{and} \quad \Sigma_{q(\beta, u)} \leftarrow (C^T R_{\text{MFVB}}^{-1} C + D_{\text{MFVB}})^{-1} \quad (12)$$

where

$$\begin{aligned} R_{\text{MFVB}} &\equiv \mu_{q(1/\sigma_\varepsilon^2)}^{-1} I, \quad D_{\text{MFVB}} \equiv \begin{bmatrix} \Sigma_\beta^{-1} & O & O \\ O & \mu_{q(1/\sigma_{\text{gbl}}^2)} I & O \\ O & O & \text{blockdiag}_{1 \leq i \leq m} \begin{bmatrix} M_{q(\Sigma^{-1})} & O \\ O & \mu_{q(1/\sigma_{\text{grp}}^2)} I \end{bmatrix} \end{bmatrix} \\ \text{and } o_{\text{MFVB}} &\equiv \begin{bmatrix} \Sigma_\beta^{-1} \mu_\beta \\ 0 \end{bmatrix}. \end{aligned} \quad (13)$$

For increasingly large numbers of groups the matrix $\Sigma_{q(\beta, u)}$ approaches a size that is untenable for random access memory storage on standard 2020s workplace computers. However, only the following relatively small sub-blocks of $\Sigma_{q(\beta, u)}$ are required for variational inference concerning the variance and covariance matrix parameters:

$$\begin{aligned} \Sigma_{q(\beta, u_{\text{gbl}})} &= \text{top left-hand } (2 + K_{\text{gbl}}) \times (2 + K_{\text{gbl}}) \text{ sub-block of } (C^T R_{\text{MFVB}}^{-1} C + D_{\text{MFVB}})^{-1}, \\ \Sigma_{q(u_{\text{lin}, i}, u_{\text{grp}, i})} &= \text{subsequent } (2 + K_{\text{grp}}) \times (2 + K_{\text{grp}}) \text{ diagonal sub-blocks of } \\ &\quad (C^T R_{\text{MFVB}}^{-1} C + D_{\text{MFVB}})^{-1} \text{ below } \Sigma_{q(\beta, u_{\text{gbl}})}, 1 \leq i \leq m, \text{ and} \\ E_q \left\{ \left(\begin{bmatrix} \beta \\ u_{\text{gbl}} \end{bmatrix} - \mu_{q(\beta, u_{\text{gbl}})} \right) \left(\begin{bmatrix} u_{\text{lin}, i} \\ u_{\text{grp}, i} \end{bmatrix} - \mu_{q(u_{\text{lin}, i}, u_{\text{grp}, i})} \right)^T \right\} &= \text{subsequent} \\ &\quad (2 + K_{\text{gbl}}) \times (2 + K_{\text{grp}}) \text{ sub-blocks of } (C^T R_{\text{MFVB}}^{-1} C + D_{\text{MFVB}})^{-1} \\ &\quad \text{to the right of } \Sigma_{q(\beta, u_{\text{gbl}})}, 1 \leq i \leq m. \end{aligned} \quad (14)$$

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

Result 2. *The mean field variational Bayes updates of $\mu_{q(\beta, u)}$ and each of the sub-blocks of $\Sigma_{q(\beta, u)}$ in (14) are expressible as a two-level sparse matrix least squares problem of the form:*

$$\|b - B\mu_{q(\beta, u)}\|^2$$

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

$$\mathbf{b}_i \equiv \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{y}_i \\ m^{-1/2} \Sigma_\beta^{-1/2} \boldsymbol{\mu}_\beta \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_i \equiv \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_i & \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{gbl},i} \\ m^{-1/2} \Sigma_\beta^{-1/2} & \mathbf{O} \\ \mathbf{O} & m^{-1/2} \mu_{q(1/\sigma_{\text{gbl}}^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

and

$$\dot{\mathbf{B}}_i \equiv \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_i & \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{grp},i} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{M}_{q(\Sigma^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & \mu_{q(1/\sigma_{\text{grp}}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}} \end{bmatrix}$$

with each of these matrices having $\tilde{n}_i = n_i + 2 + K_{\text{gbl}} + 2 + K_{\text{grp}}$ rows and with \mathbf{B}_i having $p = 2 + K_{\text{gbl}}$ columns and $\dot{\mathbf{B}}_i$ having $q = 2 + K_{\text{grp}}$ columns. The solutions are

$$\boldsymbol{\mu}_{q(\beta, \mathbf{u}_{\text{gbl}})} = \mathbf{x}_1, \quad \Sigma_{q(\beta, \mathbf{u}_{\text{gbl}})} = \mathbf{A}^{11},$$

$$\boldsymbol{\mu}_{q(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} = \mathbf{x}_{2,i}, \quad \Sigma_{q(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} = \mathbf{A}^{22,i},$$

and

$$E_q \left\{ \begin{bmatrix} \beta - \boldsymbol{\mu}_{q(\beta)} \\ \mathbf{u}_{\text{gbl}} - \boldsymbol{\mu}_{q(\mathbf{u}_{\text{gbl}})} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\text{lin},i} - \boldsymbol{\mu}_{q(\mathbf{u}_{\text{lin},i})} \\ \mathbf{u}_{\text{grp},i} - \boldsymbol{\mu}_{q(\mathbf{u}_{\text{grp},i})} \end{bmatrix}^T \right\} = \mathbf{A}^{12,i}, 1 \leq i \leq m.$$

Algorithm 2 utilizes Result 2 to facilitate streamlined computation of the variational parameters.

Lastly, we note that Algorithm 2 is loosely related to Algorithm 2 of Lee & Wand (2016). One difference is that we are treating the Gaussian, rather than Bernoulli, response situation here. In addition, we are using the recent sparse multilevel matrix results of Nolan & Wand (2018) which are amenable to higher level extensions, such as the three-level group specific curve model treated in Section 3.

2.3 Contrast Function Extension

In many curve-type data applications the data can be categorized as being from two or more types. Of particular interest in such circumstances are contrast function estimates and accompanying standard errors. The streamlined approaches used in Algorithms 1 and 2 still apply for the contrast function extension regardless of the number of categories. The two category situation, where there is a single contrast function, is described here. The extension to higher numbers of categories is straightforward.

Suppose that the (x_{ij}, y_{ij}) pairs are from one of two categories, labeled A and B , and introduce the indicator variable data:

$$\iota_{ij}^A \equiv \begin{cases} 1 & \text{if } (x_{ij}, y_{ij}) \text{ is from category } A, \\ 0 & \text{if } (x_{ij}, y_{ij}) \text{ is from category } B. \end{cases}$$

Algorithm 2 QR-decomposition-based streamlined algorithm for obtaining mean field variational Bayes approximate posterior density functions for the parameters in the Bayesian two-level group-specific curves model (10) with product density restriction (11).

Data Inputs: $\mathbf{y}_i (n_i \times 1)$, $\mathbf{X}_i (n_i \times 2)$, $\mathbf{Z}_{\text{gbl},i} (n_i \times K_{\text{gbl}})$, $\mathbf{Z}_{\text{grp},i} (n_i \times K_{\text{grp}})$, $1 \leq i \leq m$;

Hyperparameter Inputs: $\boldsymbol{\mu}_\beta (2 \times 1)$, $\boldsymbol{\Sigma}_\beta (2 \times 2)$ symmetric and positive definite,

$s_\varepsilon, \nu_\varepsilon, s_{\text{gbl}}, \nu_{\text{gbl}}, s_\Sigma, \nu_\Sigma, s_{\text{grp}}, \nu_{\text{grp}} > 0$.

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

$$\mathbf{C}_{\text{gbl},i} \leftarrow [\mathbf{X}_i \ \mathbf{Z}_{\text{gbl},i}] \ ; \ \mathbf{C}_{\text{grp},i} \leftarrow [\mathbf{X}_i \ \mathbf{Z}_{\text{grp},i}]$$

Initialize: $\mu_{\text{q}(1/\sigma_\varepsilon^2)}, \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)}, \mu_{\text{q}(1/\sigma_{\text{grp}}^2)}, \mu_{\text{q}(1/a_\varepsilon)}, \mu_{\text{q}(1/a_{\text{gbl}})}, \mu_{\text{q}(1/a_{\text{grp}})} > 0$,

$\mathbf{M}_{\text{q}(\boldsymbol{\Sigma}^{-1})} (2 \times 2), \mathbf{M}_{\text{q}(\mathbf{A}_\Sigma^{-1})} (2 \times 2)$ both symmetric and positive definite.

$\xi_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \nu_\varepsilon + \sum_{i=1}^m n_i \ ; \ \xi_{\text{q}(\sigma_{\text{gbl}}^2)} \leftarrow \nu_{\text{gbl}} + K_{\text{gbl}} \ ; \ \xi_{\text{q}(\boldsymbol{\Sigma})} \leftarrow \nu_\Sigma + 2 + m$

$\xi_{\text{q}(\sigma_{\text{grp}}^2)} \leftarrow \nu_{\text{grp}} + m K_{\text{grp}} \ ; \ \xi_{\text{q}(a_\varepsilon)} \leftarrow \nu_\varepsilon + 1 \ ; \ \xi_{\text{q}(a_{\text{gbl}})} \leftarrow \nu_{\text{gbl}} + 1 \ ; \ \xi_{\text{q}(a_{\text{grp}})} \leftarrow \nu_{\text{grp}} + 1$

$\xi_{\text{q}(\mathbf{A}_\Sigma)} \leftarrow \nu_\Sigma + 2$

Cycle:

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

$$\mathbf{b}_i \leftarrow \begin{bmatrix} \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{y}_i \\ m^{-1/2} \boldsymbol{\Sigma}_\beta^{-1/2} \boldsymbol{\mu}_\beta \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{B}_i \leftarrow \begin{bmatrix} \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_i & \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{gbl},i} \\ m^{-1/2} \boldsymbol{\Sigma}_\beta^{-1/2} & \mathbf{O} \\ \mathbf{O} & m^{-1/2} \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{\mathbf{B}}_i \leftarrow \begin{bmatrix} \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_i & \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{grp},i} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{M}_{\text{q}(\boldsymbol{\Sigma}^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & \mu_{\text{q}(1/\sigma_{\text{grp}}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}} \end{bmatrix}$$

$\mathcal{S}_2 \leftarrow \text{SOLVETWOLEVELSPARSELEASTSQUARES}(\{(\mathbf{b}_i, \mathbf{B}_i, \dot{\mathbf{B}}_i) : 1 \leq i \leq m\})$

$\boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \leftarrow \mathbf{x}_1$ component of $\mathcal{S}_2 \ ; \ \boldsymbol{\Sigma}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \leftarrow \mathbf{A}^{11}$ component of \mathcal{S}_2

$\boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{gbl}})} \leftarrow$ last K_{gbl} rows of $\boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})}$

$\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{gbl}})} \leftarrow$ bottom-right $K_{\text{gbl}} \times K_{\text{gbl}}$ sub-block of $\boldsymbol{\Sigma}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})}$

$\lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \mu_{\text{q}(1/a_\varepsilon)} \ ; \ \Lambda_{\text{q}(\boldsymbol{\Sigma})} \leftarrow \mathbf{M}_{\text{q}(\mathbf{A}_\Sigma^{-1})} \ ; \ \lambda_{\text{q}(\sigma_{\text{grp}}^2)} \leftarrow \mu_{\text{q}(1/a_{\text{grp}})}$

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

$\boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \leftarrow \mathbf{x}_{2,i}$ component of \mathcal{S}_2

$\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \leftarrow \mathbf{A}^{22,i}$ component of \mathcal{S}_2

$\boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i})} \leftarrow$ first 2 rows of $\boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}$

$\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i})} \leftarrow$ top left 2×2 sub-block of $\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}$

$\boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{grp},i})} \leftarrow$ last K_{grp} rows of $\boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}$

$\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{grp},i})} \leftarrow$ bottom right $K_{\text{grp}} \times K_{\text{grp}}$ sub-block of $\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}$

continued on a subsequent page ...

Then penalized spline models for the global mean and deviation functions for each cate-

Algorithm 2 continued. This is a continuation of the description of this algorithm that commences on a preceding page.

$$\begin{aligned}
& E_q \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right)^T \right\} \\
& \quad \leftarrow \mathbf{A}^{12,i} \text{ component of } \mathcal{S}_2 \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + \left\| \mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} - \mathbf{C}_{\text{grp},i} \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right\|^2 \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + \text{tr}(\mathbf{C}_{\text{gbl},i}^T \mathbf{C}_{\text{gbl},i} \boldsymbol{\Sigma}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})}) + \text{tr}(\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{grp},i} \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}) \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} \\
& \quad + 2 \text{tr} \left[\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{gbl},i} E_q \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right)^T \right\} \right] \\
& \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma})} \leftarrow \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma})} + \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i})} \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i})}^T + \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i})} \\
& \lambda_{\text{q}(\sigma_{\text{grp}}^2)} \leftarrow \lambda_{\text{q}(\sigma_{\text{grp}}^2)} + \left\| \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{grp},i})} \right\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{grp},i})}) \\
& \lambda_{\text{q}(\sigma_{\text{gbl}}^2)} \leftarrow \mu_{\text{q}(1/a_{\text{gbl}})} + \left\| \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{gbl}})} \right\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{gbl}})}) \\
& \mu_{\text{q}(1/\sigma_\varepsilon^2)} \leftarrow \xi_{\text{q}(\sigma_\varepsilon)} / \lambda_{\text{q}(\sigma_\varepsilon^2)} \quad ; \quad \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)} \leftarrow \xi_{\text{q}(\sigma_{\text{gbl}}^2)} / \lambda_{\text{q}(\sigma_{\text{gbl}}^2)} \\
& \mathbf{M}_{\text{q}(\boldsymbol{\Sigma}^{-1})} \leftarrow (\xi_{\text{q}(\boldsymbol{\Sigma})} - 1) \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma})}^{-1} \quad ; \quad \mu_{\text{q}(1/\sigma_{\text{grp}}^2)} \leftarrow \xi_{\text{q}(\sigma_{\text{grp}}^2)} / \lambda_{\text{q}(\sigma_{\text{grp}}^2)} \\
& \lambda_{\text{q}(a_\varepsilon)} \leftarrow \mu_{\text{q}(1/\sigma_\varepsilon^2)} + 1/(\nu_\varepsilon s_\varepsilon^2) \quad ; \quad \mu_{\text{q}(1/a_\varepsilon)} \leftarrow \xi_{\text{q}(a_\varepsilon)} / \lambda_{\text{q}(a_\varepsilon)} \\
& \boldsymbol{\Lambda}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}})} \leftarrow \text{diag}\{\text{diagonal}(\mathbf{M}_{\text{q}(\boldsymbol{\Sigma}^{-1})})\} + \{\nu_{\boldsymbol{\Sigma}} \text{diag}(s_{\boldsymbol{\Sigma},1}^2, s_{\boldsymbol{\Sigma},2}^2)\}^{-1} \\
& \mathbf{M}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}}^{-1})} \leftarrow \xi_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}})} \boldsymbol{\Lambda}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}})}^{-1} \\
& \lambda_{\text{q}(a_{\text{gbl}})} \leftarrow \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)} + 1/(\nu_{\text{gbl}} s_{\text{gbl}}^2) \quad ; \quad \mu_{\text{q}(1/a_{\text{gbl}})} \leftarrow \xi_{\text{q}(a_{\text{gbl}})} / \lambda_{\text{q}(a_{\text{gbl}})} \\
& \lambda_{\text{q}(a_{\text{grp}})} \leftarrow \mu_{\text{q}(1/\sigma_{\text{grp}}^2)} + 1/(\nu_{\text{grp}} s_{\text{grp}}^2) \quad ; \quad \mu_{\text{q}(1/a_{\text{grp}})} \leftarrow \xi_{\text{q}(a_{\text{grp}})} / \lambda_{\text{q}(a_{\text{grp}})}
\end{aligned}$$

until the increase in $\mathbf{p}(\mathbf{y}; \mathbf{q})$ is negligible.

Outputs: $\boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})}, \boldsymbol{\Sigma}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})}, \left\{ \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}, \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right\}$,

$$\begin{aligned}
& E_q \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right)^T \right\} : 1 \leq i \leq m \}, \\
& \xi_{\text{q}(\sigma_\varepsilon)}, \lambda_{\text{q}(\sigma_\varepsilon^2)}, \xi_{\text{q}(\sigma_{\text{gbl}}^2)}, \lambda_{\text{q}(\sigma_{\text{gbl}}^2)}, \xi_{\text{q}(\boldsymbol{\Sigma})}, \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma})}^{-1}, \xi_{\text{q}(\sigma_{\text{grp}}^2)}, \lambda_{\text{q}(\sigma_{\text{grp}}^2)}.
\end{aligned}$$

gory are

$$\left. \begin{aligned}
f^A(x) &= \beta_0^A + \beta_1^A x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k}^A z_{\text{gbl},k}(x) \\
g_i^A(x) &= u_{\text{lin},i0}^A + u_{\text{lin},i1}^A x + \sum_{k=1}^{K_{\text{grp}}} u_{\text{grp},ik}^A z_{\text{grp},k}(x)
\end{aligned} \right\} \text{ for category A}$$

and

$$\left. \begin{aligned}
f^B(x) &= \beta_0^A + \beta_0^{\text{BvsA}} + (\beta_1^A + \beta_1^{\text{BvsA}}) x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k}^B z_{\text{gbl},k}(x) \\
g_i^B(x) &= u_{\text{lin},i0}^B + u_{\text{lin},i1}^B x + \sum_{k=1}^{K_{\text{grp}}} u_{\text{grp},ik}^B z_{\text{grp},k}(x)
\end{aligned} \right\} \text{ for category B.}$$

This allows us to estimate the global contrast function

$$c(x) \equiv f^B(x) - f^A(x) = \beta_0^{\text{BvsA}} + \beta_1^{\text{BvsA}} x + \sum_{k=1}^{K_{\text{gbl}}} (u_{\text{gbl},k}^B - u_{\text{gbl},k}^A) z_{\text{gbl},k}(x). \quad (15)$$

The distributions on the random coefficients are

$$[u_{\text{lin},i0}^A \ u_{\text{lin},i1}^A \ u_{\text{lin},i0}^B \ u_{\text{lin},i1}^B]^T \stackrel{\text{ind.}}{\sim} N(\mathbf{0}, \Sigma)$$

and

$$u_{\text{gbl},k}^A \stackrel{\text{ind.}}{\sim} N(0, (\sigma_{\text{gbl}}^A)^2), \quad u_{\text{gbl},k}^B \stackrel{\text{ind.}}{\sim} N(0, (\sigma_{\text{gbl}}^B)^2), \quad u_{\text{grp},ik}^A \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{grp}}^2) \quad \text{and} \quad u_{\text{grp},ik}^B \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{grp}}^2)$$

independently of each other. In this two-category extension, the matrix Σ is an unstructured 4×4 covariance matrix.

Algorithms 1 and 2 can be used to achieve streamlined fitting and inference for the contrast curve extension, but with key matrices having new definitions. Firstly, the \mathbf{X}_i , $\mathbf{Z}_{\text{gbl},i}$ and $\mathbf{Z}_{\text{grp},i}$ matrices need to become:

$$\mathbf{X}_i = [\mathbf{1} \quad \mathbf{x}_i \quad \mathbf{1} - \boldsymbol{\iota}_i^A \quad (\mathbf{1} - \boldsymbol{\iota}_i^A) \odot \mathbf{x}_i],$$

$$\mathbf{Z}_{\text{gbl},i} = [\boldsymbol{\iota}_i^A \odot z_{\text{gbl},1}(\mathbf{x}_i) \cdots \boldsymbol{\iota}_i^A \odot z_{\text{gbl},K_{\text{gbl}}}(\mathbf{x}_i) \quad (\mathbf{1} - \boldsymbol{\iota}_i^A) \odot z_{\text{gbl},1}(\mathbf{x}_i) \cdots (\mathbf{1} - \boldsymbol{\iota}_i^A) \odot z_{\text{gbl},K_{\text{gbl}}}(\mathbf{x}_i)]$$

and

$$\mathbf{Z}_{\text{grp},i} = [\boldsymbol{\iota}_i^A \odot z_{\text{grp},1}(\mathbf{x}_i) \cdots \boldsymbol{\iota}_i^A \odot z_{\text{grp},K_{\text{grp}}}(\mathbf{x}_i) \quad (\mathbf{1} - \boldsymbol{\iota}_i^A) \odot z_{\text{grp},1}(\mathbf{x}_i) \cdots (\mathbf{1} - \boldsymbol{\iota}_i^A) \odot z_{\text{grp},K_{\text{grp}}}(\mathbf{x}_i)]$$

where $\boldsymbol{\iota}_i^A$ is the $n_i \times 1$ vector of ι_{ij}^A values. In the case of best linear unbiased prediction the updates for the \mathbf{B}_i and $\dot{\mathbf{B}}_i$ matrices in Algorithm 1 need to be replaced by:

$$\mathbf{B}_i \leftarrow \begin{bmatrix} \sigma_{\varepsilon}^{-1} \mathbf{X}_i & \sigma_{\varepsilon}^{-1} \mathbf{Z}_{\text{gbl},i} \\ \mathbf{O} & m^{-1/2} \begin{bmatrix} (\sigma_{\text{gbl}}^A)^{-1} \mathbf{I}_{K_{\text{gbl}}} & \mathbf{0} \\ \mathbf{0} & (\sigma_{\text{gbl}}^B)^{-1} \mathbf{I}_{K_{\text{gbl}}} \end{bmatrix} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \text{and} \quad \dot{\mathbf{B}}_i \leftarrow \begin{bmatrix} \sigma_{\varepsilon}^{-1} \mathbf{X}_i & \sigma_{\varepsilon}^{-1} \mathbf{Z}_{\text{grp},i} \\ \mathbf{O} & \mathbf{O} \\ \Sigma^{-1/2} & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp}}^{-1} \mathbf{I}_{2K_{\text{grp}}} \end{bmatrix}$$

and the output coefficient vectors change to

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}}^A \\ \hat{\mathbf{u}}_{\text{gbl}}^B \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^A \\ \hat{\mathbf{u}}_{\text{lin},i}^B \\ \hat{\mathbf{u}}_{\text{grp},i}^A \\ \hat{\mathbf{u}}_{\text{grp},i}^B \end{bmatrix}.$$

In the case of mean field variational Bayes the updates of the \mathbf{B}_i and $\dot{\mathbf{B}}_i$ matrices in Algorithm 2 need to be replaced by:

$$\mathbf{B}_i \leftarrow \begin{bmatrix} \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{X}_i & \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{Z}_{\text{gbl},i} \\ m^{-1/2} \Sigma_{\boldsymbol{\beta}}^{-1/2} & \mathbf{O} \\ \mathbf{O} & m^{-1/2} \begin{bmatrix} \mu_{q(1/(\sigma_{\text{gbl}}^A)^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/(\sigma_{\text{gbl}}^B)^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \end{bmatrix} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

and

$$\dot{\mathbf{B}}_i \leftarrow \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_i & \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{grp},i} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{M}_{q(\Sigma^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & \mu_{q(1/\sigma_{\text{grp}}^2)}^{1/2} \mathbf{I}_{2K_{\text{grp}}} \end{bmatrix}.$$

A contrast curves adjustment to the mean field variational Bayes updates is also required for some of the covariance matrix parameters. However, these calculations are comparatively simple and analogous to those given in Section S.5.

We demonstrate the use of Algorithm 2 in this setting for data from a longitudinal study on adolescent somatic growth. More detail on this data can be found in Pratt *et al.* (1989). The variables of interest are

$$\begin{aligned} y_{ij} &= j\text{th height measurement (centimetres) of subject } i, \text{ and} \\ x_{ij} &= \text{age (years) of subject } i \text{ when } y_{ij} \text{ is recorded,} \end{aligned}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n_i$. The subjects are categorized into black ethnicity and white ethnicity and comparison of mean height between the two populations is of interest. Algorithm 2 is seen to have good agreement with the data in each sub-panel of the top two plots in Figure 2. The bottom panels of Figure 2 show the estimated height gap between black and white adolescents as a function of age. For the females, there is a significant height difference only at 16-17 years old. Between 5 and 15 years, there is no obvious height difference. For the males, it is highest and (marginally) statistically significant up to about 14 years of age, peaking at 13 years of age. Between 17 and 20 years old there is no discernible height difference between the two populations.

3 Three-Level Models

The three-level version of group-specific curve models corresponds to curve-type data having two nested groupings. For example, the data in each panel of Figure 1 are first grouped according to slice, which is the level 2 group, and the slices are grouped according to tumor which is the level 3 group. We denote predictor/response pairs as (x_{ijk}, y_{ijk}) where x_{ijk} is the k th value of the predictor variable in the i th level 3 group and (i, j) th level 2 group and y_{ijk} is the corresponding value of the response variable. We let m denote the number of level 3 groups, n_i denote that number of level 2 groups in the i th level 3 group and o_{ij} denote the the number of units within the (i, j) th level 2 group. The Figure 1 data, which happen to be balanced, are such that

$$m = \text{number of tumors} = 10,$$

$$n_i = \text{number of slices for the } i\text{th tumor} = 5$$

$$\text{and } o_{ij} = \text{number of predictor/response pairs for the } i\text{th tumor and } j\text{th slice} = 128.$$

The Gaussian response three-level group specific curve model for such data is

$$\begin{aligned} y_{ijk} &= f(x_{ijk}) + g_i(x_{ijk}) + h_{ij}(x_{ijk}) + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2), \\ 1 &\leq i \leq m, \quad 1 \leq j \leq n_i, \quad 1 \leq k \leq o_{ij}, \end{aligned} \tag{16}$$

where the smooth function f is the global mean function, the g_i functions, $1 \leq i \leq m$, allow for group-specific deviations according to membership of the i th level 3 group and

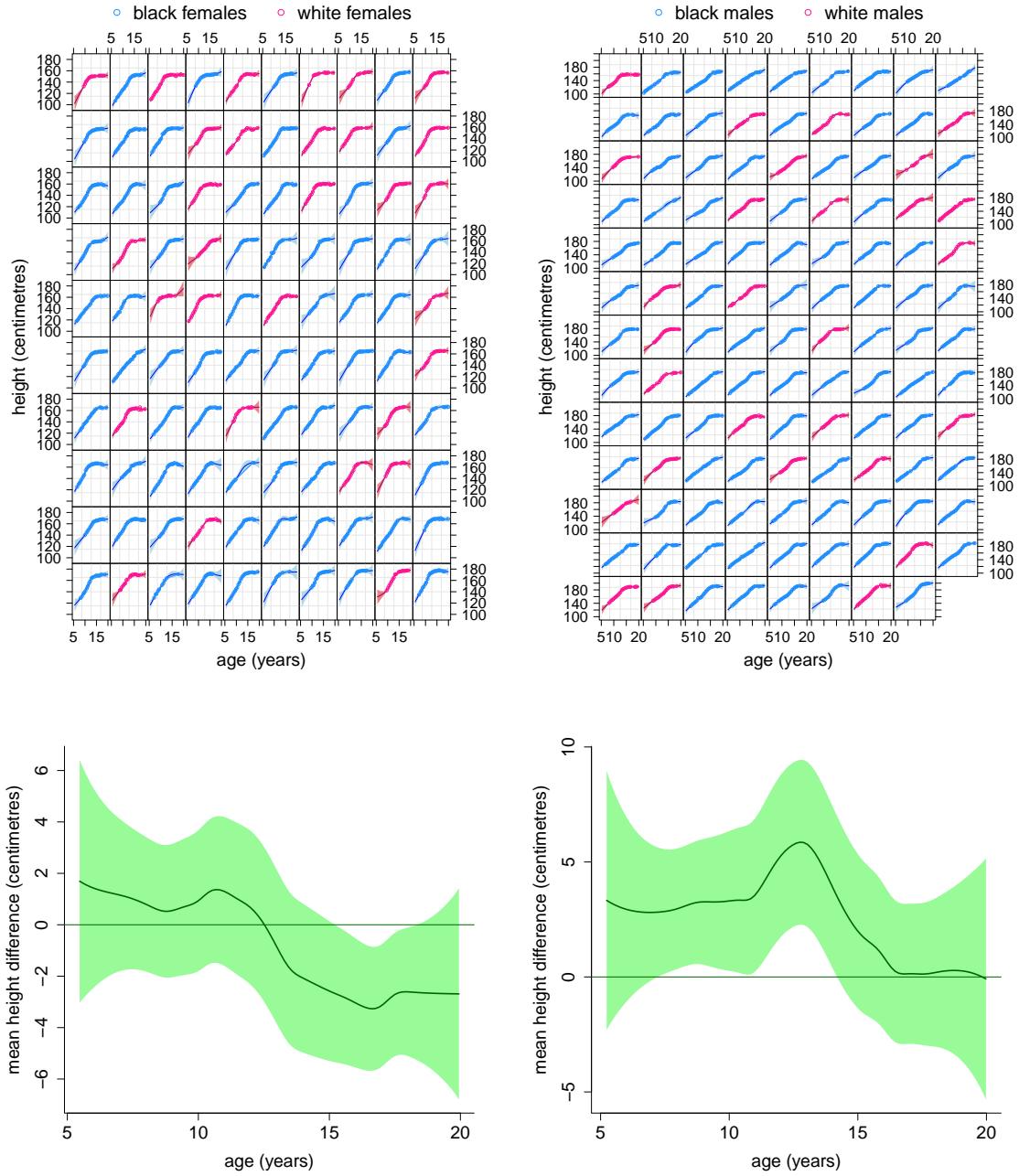


Figure 2: Top panels: fitted group-specific curves for 100 female subjects (left) and 116 male subjects (right) from the data on adolescent somatic growth (Pratt et al. 1989). The shading corresponds to approximate pointwise 99% credible intervals. Bottom panels: similar to the top panels but for the estimated contrast curve. The shaded regions correspond to approximate pointwise 95% credible intervals.

the h_{ij} , $1 \leq i \leq m$ and $1 \leq j \leq n_i$ allow for an additional level of group-specific deviations according to membership of the j th level 2 group within the i th level 3 group. The mixed model-based penalized spline models for these functions are

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K_{\text{gbl}}} u_{\text{gbl},k} z_{\text{gbl},k}(x), \quad u_{\text{gbl},k} \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{gbl}}^2),$$

$$g_i(x) = u_{\text{lin},i0}^g + u_{\text{lin},i1}^g x + \sum_{k=1}^{K_{\text{grp}}^g} u_{\text{grp},ik}^g z_{\text{grp},k}^g(x), \quad \begin{bmatrix} u_{\text{lin},i0}^g \\ u_{\text{lin},i1}^g \end{bmatrix} \stackrel{\text{ind.}}{\sim} N(\mathbf{0}, \Sigma_g), \quad u_{\text{grp},ik}^g \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{grp},g}^2)$$

and

$$h_{ij}(x) = u_{\text{lin},ij0}^h + u_{\text{lin},ij1}^h x + \sum_{k=1}^{K_{\text{grp}}^h} u_{\text{grp},ijk}^h z_{\text{grp},k}^h(x), \quad \begin{bmatrix} u_{\text{lin},ij0}^h \\ u_{\text{lin},ij1}^h \end{bmatrix} \stackrel{\text{ind.}}{\sim} N(\mathbf{0}, \Sigma_h), \quad u_{\text{grp},ijk}^h \stackrel{\text{ind.}}{\sim} N(0, \sigma_{\text{grp},h}^2),$$

with all random effect distributions independent of each other. For this three-level case we have three bases:

$$\{z_{\text{gbl},k}(\cdot) : 1 \leq k \leq K_{\text{gbl}}\}, \quad \{z_{\text{grp},k}^g(\cdot) : 1 \leq k \leq K_{\text{grp}}^g\} \quad \text{and} \quad \{z_{\text{grp},k}^h(\cdot) : 1 \leq k \leq K_{\text{grp}}^h\}.$$

The variance and covariance matrix parameters are analogous to the two-level model. For example, Σ_g and Σ_h are both unstructured 2×2 matrices corresponding to the linear components of the g_i and h_{ij} respectively.

The following notation is useful for setting up the required design matrices: if M_1, \dots, M_d is a set of matrices each having the same number of columns then

$$\text{stack}(M_i)_{1 \leq i \leq d} \equiv \begin{bmatrix} M_1 \\ \vdots \\ M_d \end{bmatrix}.$$

We then define, for $1 \leq i \leq m$ and $1 \leq j \leq n_i$,

$$\mathbf{x}_i \equiv \text{stack}_{1 \leq j \leq n_i}(\mathbf{x}_{ij}) \quad \text{and} \quad \mathbf{x}_{ij} \equiv \text{stack}_{1 \leq k \leq o_{ij}}(x_{ijk}).$$

3.1 Best Linear Unbiased Prediction

Model (16) is expressible as a Gaussian response linear mixed model as follows:

$$\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2 \mathbf{I}), \quad \mathbf{u} \sim N(\mathbf{0}, \mathbf{G}), \quad (17)$$

where the design matrices are

$$\mathbf{X} = \text{stack}_{1 \leq i \leq m}(\mathbf{X}_i) \quad \text{with} \quad \mathbf{X}_i = \text{stack}_{1 \leq j \leq n_i}(\mathbf{X}_{ij}) \quad \text{and} \quad \mathbf{X}_{ij} \equiv [\mathbf{1} \quad \mathbf{x}_{ij}]$$

and

$$\mathbf{Z} \equiv \left[\mathbf{Z}_{\text{gbl}} \quad \text{blockdiag}_{1 \leq i \leq m} \left[\text{stack}_{1 \leq j \leq n_i}([\mathbf{X}_{ij} \quad \mathbf{Z}_{\text{grp},ij}^g]) \quad \text{blockdiag}_{1 \leq j \leq n_i}([\mathbf{X}_{ij} \quad \mathbf{Z}_{\text{grp},ij}^h]) \right] \right].$$

where

$$\mathbf{Z}_{\text{gbl}} \equiv \text{stack}_{1 \leq i \leq m} \left(\text{stack}_{1 \leq j \leq n_i}(\mathbf{Z}_{\text{gbl},ij}) \right)$$

and the matrices $\mathbf{Z}_{\text{gbl},ij}$, $\mathbf{Z}_{\text{grp},ij}^g$ and $\mathbf{Z}_{\text{grp},ij}^h$, $1 \leq i \leq m$, $1 \leq j \leq n_i$, contain, respectively, spline basis functions for the global mean function f , the i th level one group deviation functions g_i and (i, j) th level two group deviation functions h_{ij} . Specifically,

$$\mathbf{Z}_{\text{gbl},ij} \equiv [z_{\text{gbl},1}(\mathbf{x}_{ij}) \cdots z_{\text{gbl},K_{\text{gbl}}}(\mathbf{x}_{ij})], \quad \mathbf{Z}_{\text{grp},ij}^g = [z_{\text{grp},1}^g(\mathbf{x}_{ij}) \cdots z_{\text{grp},K_{\text{grp}}^g}^g(\mathbf{x}_{ij})]$$

$$\text{and } \mathbf{Z}_{\text{grp},ij}^h \equiv [z_{\text{grp},1}^h(\mathbf{x}_{ij}) \cdots z_{\text{grp},K_{\text{grp}}^h}^h(\mathbf{x}_{ij})] \quad \text{for } 1 \leq i \leq m \text{ and } 1 \leq j \leq n_i.$$

The fixed and random effects vectors are

$$\beta \equiv \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{u} \equiv \begin{bmatrix} \mathbf{u}_{\text{gbl}} \\ \text{stack}_{1 \leq i \leq m} \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \\ \mathbf{u}_{\text{lin},i1}^h \\ \mathbf{u}_{\text{grp},i1}^h \\ \vdots \\ \mathbf{u}_{\text{lin},in_i}^h \\ \mathbf{u}_{\text{grp},in_i}^h \end{bmatrix} \right) \end{bmatrix} \quad \text{where} \quad \mathbf{u}_{\text{lin},i}^g \equiv \begin{bmatrix} u_{\text{lin},i0}^g \\ u_{\text{lin},i1}^g \end{bmatrix}$$

with $\mathbf{u}_{\text{grp},i}^g$, $\mathbf{u}_{\text{lin},ij}^h$ and $\mathbf{u}_{\text{grp},ij}^h$ defined similarly and the covariance matrix of \mathbf{u} is

$$\mathbf{G} = \text{Cov}(\mathbf{u}) = \begin{bmatrix} \sigma_{\text{gbl}}^2 \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \text{blockdiag}_{1 \leq i \leq m} \begin{bmatrix} \Sigma_g & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp},g}^2 \mathbf{I} \\ \mathbf{O} & \mathbf{O} & I_{n_i} \otimes \begin{bmatrix} \Sigma_h & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp},h}^2 \mathbf{I} \end{bmatrix} \end{bmatrix} \end{bmatrix}. \quad (18)$$

We define matrices in a similar way to what is given in (5). The best linear unbiased predictor of $[\beta \ \mathbf{u}]$ and corresponding covariance matrix are as shown in (6), but, with entries as described in this section. This covariance matrix grows quadratically in both m and the n_i s, and so, storage becomes impractical for large numbers of level 2 and level 3 groups. However, only certain sub-blocks are required for the addition of pointwise confidence intervals to curve estimates. In particular, we only require the non-zero sub-blocks of the general three-level sparse matrix given in Section 3 of Nolan & Wand (2018) that correspond to $(\mathbf{C}^T \mathbf{R}_{\text{BLUP}}^{-1} \mathbf{C} + \mathbf{D}_{\text{BLUP}})^{-1}$. In the case of the three-level Gaussian response linear model, Nolan & Wand's

\mathbf{A}_{11} sub-block corresponds to a $(2 + K_{\text{gbl}}) \times (2 + K_{\text{gbl}})$ matrix $\text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right);$

$\mathbf{A}_{22,i}$ sub-block corresponds to a $(2 + K_{\text{grp}}^g) \times (2 + K_{\text{grp}}^g)$ matrix $\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \right);$

$\mathbf{A}_{12,i}$ sub-block corresponds to a $(2 + K_{\text{gbl}}) \times (2 + K_{\text{grp}}^g)$ matrix

$$E \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix}^T \right\}, \quad 1 \leq i \leq m;$$

$\mathbf{A}_{22,ij}$ sub-block corresponds to a $(2 + K_{\text{grp}}^h) \times (2 + K_{\text{grp}}^h)$ matrix $\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right);$

$\mathbf{A}_{12,ij}$ sub-block corresponds to a $(2 + K_{\text{gbl}}) \times (2 + K_{\text{grp}}^h)$ matrix

$$E \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\};$$

$\mathbf{A}_{12,i,j}$ sub-block corresponds to a $(2 + K_{\text{grp}}^g) \times (2 + K_{\text{grp}}^h)$ matrix

$$E \left\{ \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i.$$

As described in Nolan, Menictas & Wand (2019), the SOLVETHREELEVELSPARSELEASTSQUARES algorithm arises in the special case where \mathbf{x} is the minimizer of the least squares problem

given in equation (8), where \mathbf{B} has the three-level sparse form and \mathbf{b} is partitioned according to that shown in equation (7) of Nolan & Wand (2018). This algorithm can be used for fitting three-level group-specific curve models by making use of Result 3.

Result 3. Computation of $[\hat{\beta}^T \hat{\mathbf{u}}^T]^T$ and each of the sub-blocks of $\text{Cov}([\hat{\beta}^T (\hat{\mathbf{u}} - \mathbf{u})^T]^T)$ listed in (7) are expressible as the three-level sparse matrix least squares form:

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

where the non-zero sub-blocks \mathbf{B} and \mathbf{b} , according to the notation in Section 3.1 of Nolan & Wand (2018), are for $1 \leq i \leq m$ and $1 \leq j \leq n_i$:

$$\mathbf{b}_{ij} \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{y}_{ij} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_{ij} \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_{ij} & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{gbl},ij} \\ \mathbf{O} & (\sum_{i=1}^m n_i)^{-1/2} \sigma_{\text{gbl}}^{-1} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{\mathbf{B}}_{ij} \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_{ij} & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{grp},ij}^g \\ \mathbf{O} & \mathbf{O} \\ n_i^{-1/2} \Sigma_g^{-1/2} & \mathbf{O} \\ \mathbf{O} & n_i^{-1/2} \sigma_{\text{grp},g}^{-1} \mathbf{I}_{K_{\text{grp}}^g} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \text{and} \quad \ddot{\mathbf{B}}_{ij} \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_{ij} & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{grp},ij}^h \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \Sigma_h^{-1/2} & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp},h}^{-1} \mathbf{I}_{K_{\text{grp}}^h} \end{bmatrix}$$

with each of these matrices having $\tilde{o}_{ij} = o_{ij} + K_{\text{gbl}} + 2 + K_{\text{grp}}^g + 2 + K_{\text{grp}}^h$ rows and with \mathbf{B}_i having $p = 2 + K_{\text{gbl}}$ columns, $\dot{\mathbf{B}}_i$ having $q_1 = 2 + K_{\text{grp}}^g$ columns and $\ddot{\mathbf{B}}_i$ having $q_2 = 2 + K_{\text{grp}}^h$ columns. The solutions are

$$\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} \end{bmatrix} = \mathbf{x}_1, \quad \text{Cov} \left(\begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right) = \mathbf{A}^{11},$$

$$\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g \end{bmatrix} = \mathbf{x}_{2,i}, \quad E \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix}^T \right\} = \mathbf{A}^{12,i},$$

$$\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \right) = \mathbf{A}^{22,i}, \quad 1 \leq i \leq m,$$

$$\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h \end{bmatrix} = \mathbf{x}_{2,ij}, \quad E \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\} = \mathbf{A}^{12,ij},$$

$$E \left\{ \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\} = \mathbf{A}^{12,i,j}$$

and

$$\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right) = \mathbf{A}^{22,ij}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i.$$

Algorithm 3 Streamlined algorithm for obtaining best linear unbiased predictions and corresponding covariance matrix components for the two-level group specific curves model.

Inputs: $\mathbf{y}_{ij}(o_{ij} \times 1)$, $\mathbf{X}_{ij}(o_{ij} \times 2)$, $\mathbf{Z}_{\text{gbl},ij}(o_{ij} \times K_{\text{gbl}})$, $\mathbf{Z}_{\text{grp},ij}^g(o_{ij} \times K_{\text{grp}}^g)$,
 $\mathbf{Z}_{\text{grp},ij}^h(o_{ij} \times K_{\text{grp}}^h)$, $1 \leq i \leq m$, $1 \leq j \leq n_i$; $\sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \sigma_{\text{grp},g}^2, \sigma_{\text{grp},h}^2 > 0$,
 $\Sigma_g(2 \times 2)$, $\Sigma_h(2 \times 2)$, symmetric and positive definite.

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

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

$$\mathbf{b}_{ij} \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{y}_{ij} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{B}_{ij} \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_{ij} & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{gbl},ij} \\ \mathbf{O} & (\sum_{i=1}^m n_i)^{-1/2} \sigma_{\text{gbl}}^{-1} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

$$\dot{\mathbf{B}}_{ij} \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_{ij} & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{grp},ij}^g \\ \mathbf{O} & \mathbf{O} \\ n_i^{-1/2} \Sigma_g^{-1/2} & \mathbf{O} \\ \mathbf{O} & n_i^{-1/2} \sigma_{\text{grp},g}^{-1} \mathbf{I}_{K_{\text{grp}}^g} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}, \ddot{\mathbf{B}}_{ij} \leftarrow \begin{bmatrix} \sigma_\varepsilon^{-1} \mathbf{X}_{ij} & \sigma_\varepsilon^{-1} \mathbf{Z}_{\text{grp},ij}^h \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \Sigma_h^{-1/2} & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp},h}^{-1} \mathbf{I}_{K_{\text{grp}}^h} \end{bmatrix}$$

$\mathcal{S}_3 \leftarrow \text{SOLVETHREELSPARSELEASTSQUARES}(\{(\mathbf{b}_{ij}, \mathbf{B}_{ij}, \dot{\mathbf{B}}_{ij}, \ddot{\mathbf{B}}_{ij}) : 1 \leq i \leq m, \\ 1 \leq j \leq n_i\})$

$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} \end{bmatrix} \leftarrow \mathbf{x}_1 \text{ component of } \mathcal{S}_3$; $\text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right) \leftarrow \mathbf{A}^{11} \text{ component of } \mathcal{S}_3$

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

$$\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g \end{bmatrix} \leftarrow \mathbf{x}_{2,i} \text{ component of } \mathcal{S}_3$$

$$\text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \right) \leftarrow \mathbf{A}^{22,i} \text{ component of } \mathcal{S}_3$$

$$E \left\{ \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix}^T \right\} \leftarrow \mathbf{A}^{12,i} \text{ component of } \mathcal{S}_3$$

continued on a subsequent page ...

A derivation of Result 3 is given in Section S.7 of the web-supplement. Result 3 combined with Theorem 4 of Nolan & Wand (2018) leads to Algorithm 3. The SOLVETHREELSPARSELEASTSQUARES algorithm is given in Section S.12.

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

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

$$\begin{aligned}
& \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h \end{bmatrix} \leftarrow \mathbf{x}_{2,ij} \text{ component of } \mathcal{S}_3 \\
& E \left\{ \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\} \leftarrow \mathbf{A}^{12,ij} \text{ component of } \mathcal{S}_3 \\
& E \left\{ \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\} \leftarrow \mathbf{A}^{12,i,j} \text{ component of } \mathcal{S}_3 \\
& \text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right) \leftarrow \mathbf{A}^{22,ij} \text{ component of } \mathcal{S}_3
\end{aligned}$$

Output:

$$\begin{aligned}
& \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} \end{bmatrix}, \text{Cov} \left(\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \right), \left\{ \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} \end{bmatrix}, \text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix} \right), \right. \right. \\
& \quad \left. \left. E \left\{ \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i} - \mathbf{u}_{\text{lin},i} \\ \hat{\mathbf{u}}_{\text{grp},i} - \mathbf{u}_{\text{grp},i} \end{bmatrix}^T \right\} \right) : 1 \leq i \leq m, \right. \\
& \quad \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h \end{bmatrix}, E \left\{ \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}}_{\text{gbl}} - \mathbf{u}_{\text{gbl}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\}, \right. \\
& \quad \left. E \left\{ \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},i}^g - \mathbf{u}_{\text{lin},i}^g \\ \hat{\mathbf{u}}_{\text{grp},i}^g - \mathbf{u}_{\text{grp},i}^g \end{bmatrix} \begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix}^T \right\}, \text{Cov} \left(\begin{bmatrix} \hat{\mathbf{u}}_{\text{lin},ij}^h - \mathbf{u}_{\text{lin},ij}^h \\ \hat{\mathbf{u}}_{\text{grp},ij}^h - \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right) \right) : \\
& \quad \left. 1 \leq i \leq m, 1 \leq j \leq n_i \right\}
\end{aligned}$$

3.2 Mean Field Variational Bayes

A Bayesian extension of (17) and (18) is:

$$\begin{aligned}
& \mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \sigma_\varepsilon^2 \mathbf{I}), \quad \mathbf{u} | \sigma_{\text{gbl}}^2, \sigma_{\text{grp},g}^2, \boldsymbol{\Sigma}_g, \sigma_{\text{grp},h}^2, \boldsymbol{\Sigma}_h \sim N(\mathbf{0}, \mathbf{G}), \quad \mathbf{G} \text{ as defined in (18),} \\
& \boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta), \quad \sigma_\varepsilon^2 | a_\varepsilon \sim \text{Inverse-}\chi^2(\nu_\varepsilon, 1/a_\varepsilon), \quad a_\varepsilon \sim \text{Inverse-}\chi^2(1, 1/(\nu_\varepsilon s_\varepsilon^2)), \\
& \sigma_{\text{gbl}}^2 | a_{\text{gbl}} \sim \text{Inverse-}\chi^2(\nu_{\text{gbl}}, 1/a_{\text{gbl}}), \quad a_{\text{gbl}} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\text{gbl}} s_{\text{gbl}}^2)), \\
& \sigma_{\text{grp},g}^2 | a_{\text{grp},g} \sim \text{Inverse-}\chi^2(\nu_{\text{grp},g}, 1/a_{\text{grp},g}), \quad a_{\text{grp},g} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\text{grp},g} s_{\text{grp},g}^2)), \\
& \sigma_{\text{grp},h}^2 | a_{\text{grp},h} \sim \text{Inverse-}\chi^2(\nu_{\text{grp},h}, 1/a_{\text{grp},h}), \quad a_{\text{grp},h} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\text{grp},h} s_{\text{grp},h}^2)), \\
& \boldsymbol{\Sigma}_g | \mathbf{A}_{\boldsymbol{\Sigma}_g} \sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}_g} + 2, \mathbf{A}_{\boldsymbol{\Sigma}_g}^{-1}), \\
& \mathbf{A}_{\boldsymbol{\Sigma}_g} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{\mathbf{A}_{\boldsymbol{\Sigma}_g}}), \quad \boldsymbol{\Lambda}_{\mathbf{A}_{\boldsymbol{\Sigma}_g}} \equiv \{\nu_{\boldsymbol{\Sigma}_g} \text{diag}(s_{\boldsymbol{\Sigma}_g,1}^2, s_{\boldsymbol{\Sigma}_g,2}^2)\}^{-1}, \\
& \boldsymbol{\Sigma}_h | \mathbf{A}_{\boldsymbol{\Sigma}_h} \sim \text{Inverse-G-Wishart}(G_{\text{full}}, \nu_{\boldsymbol{\Sigma}_h} + 2, \mathbf{A}_{\boldsymbol{\Sigma}_h}^{-1}), \\
& \mathbf{A}_{\boldsymbol{\Sigma}_h} \sim \text{Inverse-G-Wishart}(G_{\text{diag}}, 1, \boldsymbol{\Lambda}_{\mathbf{A}_{\boldsymbol{\Sigma}_h}}), \quad \boldsymbol{\Lambda}_{\mathbf{A}_{\boldsymbol{\Sigma}_h}} \equiv \{\nu_{\boldsymbol{\Sigma}_h} \text{diag}(s_{\boldsymbol{\Sigma}_h,1}^2, s_{\boldsymbol{\Sigma}_h,2}^2)\}^{-1}.
\end{aligned} \tag{19}$$

The following mean field restriction is imposed on the joint posterior density function of all parameters in (19):

$$\begin{aligned} p(\boldsymbol{\beta}, \mathbf{u}, a_\varepsilon, a_{\text{gbl}}, a_{\text{grp},g}, \mathbf{A}_{\boldsymbol{\Sigma}_g}, a_{\text{grp},h}, \mathbf{A}_{\boldsymbol{\Sigma}_h}, \sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \sigma_{\text{grp},g}^2, \boldsymbol{\Sigma}_g, \sigma_{\text{grp},h}^2, \boldsymbol{\Sigma}_h | \mathbf{y}) \\ \approx q(\boldsymbol{\beta}, \mathbf{u}, a_\varepsilon, a_{\text{gbl}}, a_{\text{grp},g}, \mathbf{A}_{\boldsymbol{\Sigma}_g}, a_{\text{grp},h}, \mathbf{A}_{\boldsymbol{\Sigma}_h}) q(\sigma_\varepsilon^2, \sigma_{\text{gbl}}^2, \sigma_{\text{grp},g}^2, \boldsymbol{\Sigma}_g, \sigma_{\text{grp},h}^2, \boldsymbol{\Sigma}_h). \end{aligned} \quad (20)$$

The optimal q -density functions for the parameters of interest are

$q^*(\boldsymbol{\beta}, \mathbf{u})$ has a $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta}, \mathbf{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \mathbf{u})})$ distribution,

$q^*(\sigma_\varepsilon^2)$ has an Inverse- $\chi^2(\xi_{q(\sigma_\varepsilon^2)}, \lambda_{q(\sigma_\varepsilon^2)})$ distribution,

$q^*(\sigma_{\text{gbl}}^2)$ has an Inverse- $\chi^2(\xi_{q(\sigma_{\text{gbl}}^2)}, \lambda_{q(\sigma_{\text{gbl}}^2)})$ distribution,

$q^*(\sigma_{\text{grp},g}^2)$ has an Inverse- $\chi^2(\xi_{q(\sigma_{\text{grp},g}^2)}, \lambda_{q(\sigma_{\text{grp},g}^2)})$ distribution

$q^*(\sigma_{\text{grp},h}^2)$ has an Inverse- $\chi^2(\xi_{q(\sigma_{\text{grp},h}^2)}, \lambda_{q(\sigma_{\text{grp},h}^2)})$ distribution

$q^*(\boldsymbol{\Sigma}_g)$ has an Inverse-G-Wishart($G_{\text{full}}, \xi_{q(\boldsymbol{\Sigma}_g)}, \mathbf{\Lambda}_{q(\boldsymbol{\Sigma}_g)}$) distribution

and $q^*(\boldsymbol{\Sigma}_h)$ has an Inverse-G-Wishart($G_{\text{full}}, \xi_{q(\boldsymbol{\Sigma}_h)}, \mathbf{\Lambda}_{q(\boldsymbol{\Sigma}_h)}$) distribution.

The optimal q -density parameters are determined through an iterative coordinate ascent algorithm, details of which are given in Section S.10 of the web-supplement. As in the two-level case, the updates for $\boldsymbol{\mu}_{q(\boldsymbol{\beta}, \mathbf{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \mathbf{u})}$ may be written in the same form as (12) but with a three-level version of the \mathbf{C} matrix and

$$D_{\text{MFVB}} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\mu}_{q(1/\sigma_{\text{gbl}}^2)} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{I}_m \otimes \begin{bmatrix} M_{q(\boldsymbol{\Sigma}_g^{-1})} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\mu}_{q(1/\sigma_{\text{grp},g}^2)} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{I}_{n_i} \otimes \begin{bmatrix} M_{q(\boldsymbol{\Sigma}_h^{-1})} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\mu}_{q(1/\sigma_{\text{grp},h}^2)} \mathbf{I} \end{bmatrix} \end{bmatrix} \end{bmatrix}. \quad (21)$$

For large numbers of level 2 and level 3 groups, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \mathbf{u})}$'s size becomes infeasible to deal with. However, only relatively small sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta}, \mathbf{u})}$ are needed for variational inference regarding the variance and covariance parameters. These sub-block positions correspond to the non-zero sub-block positions of a general three-level sparse matrix defined

in Section 3 of Nolan & Wand (2018). Here, Nolan & Wand's

\mathbf{A}_{11} sub-block corresponds to a $(2 + K_{\text{gbl}}) \times (2 + K_{\text{gbl}})$ matrix $\Sigma_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})}$;

$\mathbf{A}_{22,i}$ sub-block corresponds to a $(2 + K_{\text{grp}}^g) \times (2 + K_{\text{grp}}^g)$ matrix $\Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)}$;

$\mathbf{A}_{12,i}$ sub-block corresponds to a $(2 + K_{\text{gbl}}) \times (2 + K_{\text{grp}}^g)$ matrix

$$E \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right)^T \right\}, \quad 1 \leq i \leq m;$$

$\mathbf{A}_{22,ij}$ sub-block corresponds to a $(2 + K_{\text{grp}}^h) \times (2 + K_{\text{grp}}^h)$ matrix $\Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)}$;

$\mathbf{A}_{12,ij}$ sub-block corresponds to a $(2 + K_{\text{gbl}}) \times (2 + K_{\text{grp}}^h)$ matrix

$$E \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\};$$

$\mathbf{A}_{12,i,j}$ sub-block corresponds to a $(2 + K_{\text{grp}}^g) \times (2 + K_{\text{grp}}^h)$ matrix

$$E \left\{ \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\},$$

$1 \leq i \leq m, 1 \leq j \leq n_i.$

(22)

We appeal to Result 4 for a streamlined mean field variational Bayes algorithm.

Result 4. *The mean field variational Bayes updates of $\mu_{\mathbf{q}(\beta, \mathbf{u})}$ and each of the sub-blocks of $\Sigma_{\mathbf{q}(\beta, \mathbf{u})}$ in (22) are expressible as a three-level sparse matrix least squares problem of the form:*

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

where the non-zero sub-blocks \mathbf{B} and \mathbf{b} , according to the notation in Section 3.1 of Nolan & Wand (2018), are for $1 \leq i \leq m$ and $1 \leq j \leq n_i$.

$$\mathbf{b}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{y}_{ij} \\ (\sum_{i=1}^m n_i)^{-1/2} \Sigma_{\beta}^{-1/2} \mu_{\beta} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{Z}_{\text{gbl},ij} \\ (\sum_{i=1}^m n_i)^{-1/2} \Sigma_{\beta}^{-1/2} & \mathbf{O} \\ \mathbf{O} & (\sum_{i=1}^m n_i)^{-1/2} \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{\mathbf{B}}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{Z}_{\text{grp},ij}^g \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ n_i^{-1/2} \mathbf{M}_{\mathbf{q}(\Sigma_g^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & n_i^{-1/2} \mu_{\mathbf{q}(1/\sigma_{\text{grp},g}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}^g} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \text{and} \quad \ddot{\mathbf{B}}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{Z}_{\text{grp},ij}^h \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{M}_{\mathbf{q}(\Sigma_h^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & \mu_{\mathbf{q}(1/\sigma_{\text{grp},h}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}^h} \end{bmatrix}$$

with each of these matrices having $\tilde{o}_{ij} = o_{ij} + 2 + K_{\text{gbl}} + 2 + K_{\text{grp}}^g + 2 + K_{\text{grp}}^h$ rows and with \mathbf{B}_i having $p = 2 + K_{\text{gbl}}$ columns, $\dot{\mathbf{B}}_i$ having $q_1 = 2 + K_{\text{grp}}^g$ columns and $\ddot{\mathbf{B}}_{ij}$ having $q_2 = 2 + K_{\text{grp}}^h$ columns. The solutions are

$$\begin{aligned}\mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} &= \mathbf{x}_1, \quad \Sigma_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} = \mathbf{A}^{11}, \\ \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} &= \mathbf{x}_{2,i}, \quad E_q \left\{ \begin{bmatrix} \beta - \mu_{\mathbf{q}(\beta)} \\ \mathbf{u}_{\text{gbl}} - \mu_{\mathbf{q}(\mathbf{u}_{\text{gbl}})} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\text{lin},i}^g - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g)} \\ \mathbf{u}_{\text{grp},i}^g - \mu_{\mathbf{q}(\mathbf{u}_{\text{grp},i}^g)} \end{bmatrix}^T \right\} = \mathbf{A}^{12,i}, \\ \Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} &= \mathbf{A}^{22,i}, \quad 1 \leq i \leq m, \\ \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} &= \mathbf{x}_{2,ij}, \quad E_q \left\{ \begin{bmatrix} \beta - \mu_{\mathbf{q}(\beta)} \\ \mathbf{u}_{\text{gbl}} - \mu_{\mathbf{q}(\mathbf{u}_{\text{gbl}})} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h)} \\ \mathbf{u}_{\text{grp},ij}^h - \mu_{\mathbf{q}(\mathbf{u}_{\text{grp},ij}^h)} \end{bmatrix}^T \right\} = \mathbf{A}^{12,ij}, \\ E_q \left\{ \begin{bmatrix} \mathbf{u}_{\text{lin},i}^g - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g)} \\ \mathbf{u}_{\text{grp},i}^g - \mu_{\mathbf{q}(\mathbf{u}_{\text{grp},i}^g)} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h)} \\ \mathbf{u}_{\text{grp},ij}^h - \mu_{\mathbf{q}(\mathbf{u}_{\text{grp},ij}^h)} \end{bmatrix}^T \right\} &= \mathbf{A}^{12,i,j}\end{aligned}$$

and

$$\Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} = \mathbf{A}^{22,ij}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n_i.$$

Algorithm 4 makes use of Result 4 to facilitate streamlined computation of all variational parameters in the three-level group specific curves model.

Figure 3 provides illustration of Algorithm 4 by showing the fits to the Figure 1 ultrasound data. Posterior mean curves and (narrow) 99% pointwise credible intervals are shown. As discussed in the next section, such fits can be obtained rapidly and accurately and Algorithm 4 is scalable to much larger data sets of the type illustrated by Figures 1 and 3.

Algorithm 4 QR-decomposition-based streamlined algorithm for obtaining mean field variational Bayes approximate posterior density functions for the parameters in the Bayesian three-level group-specific curves model (19) with product density restriction (20)

Data Inputs: $\mathbf{y}_{ij}(o_{ij} \times 1)$, $\mathbf{X}_{ij}(o_{ij} \times 2)$, $\mathbf{Z}_{\text{gbl},ij}(o_{ij} \times K_{\text{gbl}})$, $\mathbf{Z}_{\text{grp},ij}^g(o_{ij} \times K_{\text{grp}}^g)$,

$\mathbf{Z}_{\text{grp},ij}^h(o_{ij} \times K_{\text{grp}}^h)$ $1 \leq i \leq m$, $1 \leq j \leq n_i$.

Hyperparameter Inputs: $\boldsymbol{\mu}_\beta(2 \times 1)$, $\boldsymbol{\Sigma}_\beta(2 \times 2)$ symmetric and positive definite,

$s_\varepsilon, \nu_\varepsilon, s_{\text{gbl}}, \nu_{\text{gbl}}, s_{\boldsymbol{\Sigma}_g}, \nu_{\boldsymbol{\Sigma}_g}, s_{\boldsymbol{\Sigma}_g}, \nu_{\boldsymbol{\Sigma}_g}, s_{\text{grp},g}, \nu_{\text{grp},g}, s_{\boldsymbol{\Sigma}_h}, \nu_{\boldsymbol{\Sigma}_h}, s_{\text{grp},h}, \nu_{\text{grp},h} > 0$.

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

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

$\mathbf{C}_{\text{gbl},ij} \leftarrow [\mathbf{X}_{ij} \ \mathbf{Z}_{\text{gbl},ij}]$; $\mathbf{C}_{\text{grp},ij}^g \leftarrow [\mathbf{X}_{ij} \ \mathbf{Z}_{\text{grp},ij}^g]$; $\mathbf{C}_{\text{grp},ij}^h \leftarrow [\mathbf{X}_{ij} \ \mathbf{Z}_{\text{grp},ij}^h]$

Initialize: $\mu_{\text{q}(1/\sigma_\varepsilon^2)}, \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)}, \mu_{\text{q}(1/\sigma_{\text{grp},g}^2)}, \mu_{\text{q}(1/\sigma_{\text{grp},h}^2)}, \mu_{\text{q}(1/a_\varepsilon)}, \mu_{\text{q}(1/a_{\text{gbl}})},$

$\mu_{\text{q}(1/a_{\text{grp},g})}, \mu_{\text{q}(1/a_{\text{grp},h})} > 0$, $\mathbf{M}_{\text{q}(\boldsymbol{\Sigma}_g^{-1})}(2 \times 2)$, $\mathbf{M}_{\text{q}(\boldsymbol{\Sigma}_h^{-1})}(2 \times 2)$,

$\mathbf{M}_{\text{q}(\mathbf{A}_g^{-1})}(2 \times 2)$, $\mathbf{M}_{\text{q}(\mathbf{A}_h^{-1})}(2 \times 2)$ symmetric and positive definite.

$\xi_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \nu_\varepsilon + \sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}$; $\xi_{\text{q}(\sigma_{\text{gbl}}^2)} \leftarrow \nu_{\text{gbl}} + K_{\text{gbl}}$; $\xi_{\text{q}(\boldsymbol{\Sigma}_g)} \leftarrow \nu_{\boldsymbol{\Sigma}_g} + 2 + m$

$\xi_{\text{q}(\boldsymbol{\Sigma}_h)} \leftarrow \nu_{\boldsymbol{\Sigma}_h} + 2 + \sum_{i=1}^m n_i$; $\xi_{\text{q}(\sigma_{\text{grp},g}^2)} \leftarrow \nu_{\text{grp},g} + m K_{\text{grp},g}^g$

$\xi_{\text{q}(\sigma_{\text{grp},h}^2)} \leftarrow \nu_{\text{grp},h} + K_{\text{grp},h}^h \sum_{i=1}^m n_i$; $\xi_{\text{q}(a_\varepsilon)} \leftarrow \nu_\varepsilon + 1$; $\xi_{\text{q}(a_{\text{gbl}})} \leftarrow \nu_{\text{gbl}} + 1$

$\xi_{\text{q}(a_{\text{grp},g})} \leftarrow \nu_{\text{grp},g} + 1$; $\xi_{\text{q}(a_{\text{grp},h})} \leftarrow \nu_{\text{grp},h} + 1$; $\xi_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_g})} \leftarrow \nu_{\boldsymbol{\Sigma}_g} + 2$; $\xi_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_h})} \leftarrow \nu_{\boldsymbol{\Sigma}_h} + 2$

Cycle:

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

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

$$\mathbf{b}_{ij} \leftarrow \begin{bmatrix} \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{y}_{ij} \\ n_i^{-1/2} \boldsymbol{\Sigma}_\beta^{-1/2} \boldsymbol{\mu}_\beta \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \mathbf{B}_{ij} \leftarrow \begin{bmatrix} \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{gbl},ij} \\ (\sum_{i=1}^m n_i)^{-1/2} \boldsymbol{\Sigma}_\beta^{-1/2} & \mathbf{O} \\ \mathbf{O} & (\sum_{i=1}^m n_i)^{-1/2} \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{\mathbf{B}}_{ij} \leftarrow \begin{bmatrix} \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\text{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{grp},ij}^g \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ n_i^{-1/2} \mathbf{M}_{\text{q}(\boldsymbol{\Sigma}_g^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & n_i^{-1/2} \mu_{\text{q}(1/\sigma_{\text{grp},g}^2)}^{1/2} \mathbf{I}_{K_{\text{grp},g}^g} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

continued on a subsequent page ...

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

$$\ddot{B}_{ij} \leftarrow \begin{bmatrix} \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} X_i & \mu_{q(1/\sigma_\varepsilon^2)}^{1/2} Z_{\text{grp},ij}^h \\ O & O \\ O & O \\ O & O \\ O & O \\ M_{q(\Sigma_h^{-1})}^{1/2} & O \\ O & \mu_{q(1/\sigma_{\text{grp},h}^2)}^{1/2} I_{K_{\text{grp}}^h} \end{bmatrix}$$

$$S_4 \leftarrow \text{SOLVETHREELLEVELSPARSELEASTSQUARES}\left(\{(b_{ij}, B_{ij}, \dot{B}_{ij}, \ddot{B}_{ij}) : 1 \leq i \leq m, 1 \leq j \leq n_i\}\right)$$

$$\mu_{q(\beta, u_{\text{gbl}})} \leftarrow x_1 \text{ component of } S_4 \quad ; \quad \Sigma_{q(\beta, u_{\text{gbl}})} \leftarrow A^{11} \text{ component of } S_4$$

$$\mu_{q(u_{\text{gbl}})} \leftarrow \text{last } K_{\text{gbl}} \text{ rows of } \mu_{q(\beta, u_{\text{gbl}})}$$

$$\Sigma_{q(u_{\text{gbl}})} \leftarrow \text{bottom-right } K_{\text{gbl}} \times K_{\text{gbl}} \text{ sub-block of } \Sigma_{q(\beta, u_{\text{gbl}})}$$

$$\lambda_{q(\sigma_\varepsilon^2)} \leftarrow \mu_{q(1/a_\varepsilon)} \quad ; \quad \Lambda_{q(\Sigma_g)} \leftarrow M_{q(A_{\Sigma_g}^{-1})} \quad ; \quad \Lambda_{q(\Sigma_h)} \leftarrow M_{q(A_{\Sigma_h}^{-1})}$$

$$\lambda_{q(\sigma_{\text{grp},g}^2)} \leftarrow \mu_{q(1/a_{\text{grp},g})} \quad ; \quad \lambda_{q(\sigma_{\text{grp},g}^2)} \leftarrow \mu_{q(1/a_{\text{grp},g})}$$

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

$$\mu_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)} \leftarrow x_{2,i} \text{ component of } S_4$$

$$\Sigma_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)} \leftarrow A^{22,i} \text{ component of } S_4$$

$$\mu_{q(u_{\text{lin},i}^g)} \leftarrow \text{first 2 rows of } \mu_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)}$$

$$\Sigma_{q(u_{\text{lin},i}^g)} \leftarrow \text{top left } 2 \times 2 \text{ sub-block of } \Sigma_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)}$$

$$\mu_{q(u_{\text{grp},i}^g)} \leftarrow \text{last } K_{\text{grp}}^g \text{ rows of } \mu_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)}$$

$$\Sigma_{q(u_{\text{grp},i}^g)} \leftarrow \text{bottom right } K_{\text{grp}}^g \times K_{\text{grp}}^g \text{ sub-block of } \Sigma_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)}$$

$$E_q \left\{ \left(\begin{bmatrix} \beta \\ u_{\text{gbl}} \end{bmatrix} - \mu_{q(\beta, u_{\text{gbl}})} \right) \left(\begin{bmatrix} u_{\text{lin},i}^g \\ u_{\text{grp},i}^g \end{bmatrix} - \mu_{q(u_{\text{lin},i}^g, u_{\text{grp},i}^g)} \right)^T \right\} \\ \leftarrow A^{12,i} \text{ component of } S_4$$

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

$$\mu_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)} \leftarrow x_{2,ij} \text{ component of } S_4$$

$$\Sigma_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)} \leftarrow A^{22,ij} \text{ component of } S_4$$

$$\mu_{q(u_{\text{lin},ij}^h)} \leftarrow \text{first 2 rows of } \mu_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)}$$

$$\Sigma_{q(u_{\text{lin},ij}^h)} \leftarrow \text{top left } 2 \times 2 \text{ sub-block of } \Sigma_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)}$$

$$\mu_{q(u_{\text{grp},ij}^h)} \leftarrow \text{last } K_{\text{grp}}^h \text{ rows of } \mu_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)}$$

$$\Sigma_{q(u_{\text{grp},ij}^h)} \leftarrow \text{bottom right } K_{\text{grp}}^h \times K_{\text{grp}}^h \text{ sub-block of } \Sigma_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)}$$

$$E_q \left\{ \left(\begin{bmatrix} \beta \\ u_{\text{gbl}} \end{bmatrix} - \mu_{q(\beta, u_{\text{gbl}})} \right) \left(\begin{bmatrix} u_{\text{lin},ij}^h \\ u_{\text{grp},ij}^h \end{bmatrix} - \mu_{q(u_{\text{lin},ij}^h, u_{\text{grp},ij}^h)} \right)^T \right\} \\ \leftarrow A^{12,ij} \text{ component of } S_4$$

continued on a subsequent page . . .

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

$$\begin{aligned}
& E_q \left\{ \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\} \\
& \quad \leftarrow \mathbf{A}^{12,i,j} \text{ component of } \mathcal{S}_4 \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + \left\| \mathbf{y}_{ij} - \mathbf{C}_{\text{gbl},ij} \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} - \mathbf{C}_{\text{grp},ij}^g \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right. \\
& \quad \left. - \mathbf{C}_{\text{grp},ij}^h \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right\|^2 \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + \text{tr}(\mathbf{C}_{\text{gbl},ij}^T \mathbf{C}_{\text{gbl},ij} \boldsymbol{\Sigma}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})}) + \text{tr}((\mathbf{C}_{\text{grp},ij}^g)^T \mathbf{C}_{\text{grp},ij}^g \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)}) \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + \text{tr}((\mathbf{C}_{\text{grp},ij}^h)^T \mathbf{C}_{\text{grp},ij}^h \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)}) \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + 2 \text{tr} \left[\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{gbl},i} E_q \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \right. \right. \\
& \quad \left. \left. \times \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right)^T \right\} \right] \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + 2 \text{tr} \left[(\mathbf{C}_{\text{grp},ij}^g)^T \mathbf{C}_{\text{gbl},ij} E_q \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \right. \right. \\
& \quad \left. \left. \times \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\} \right] \\
& \lambda_{\text{q}(\sigma_\varepsilon^2)} \leftarrow \lambda_{\text{q}(\sigma_\varepsilon^2)} + 2 \text{tr} \left[(\mathbf{C}_{\text{grp},ij}^g)^T \mathbf{C}_{\text{grp},ij}^h E_q \left\{ \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right) \right. \right. \\
& \quad \left. \left. \times \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\} \right] \\
& \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma}_h)} \leftarrow \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma}_h)} + \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h)} \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h)}^T + \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},ij}^h)} \\
& \lambda_{\text{q}(\sigma_{\text{grp},h}^2)} \leftarrow \lambda_{\text{q}(\sigma_{\text{grp},h}^2)} + \left\| \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{grp},ij}^h)} \right\|^2 + \text{tr} \left(\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{grp},ij}^h)} \right) \\
& \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma}_g)} \leftarrow \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma}_g)} + \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}^g)} \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{lin},i}^g)}^T + \boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{lin},i}^g)} \\
& \lambda_{\text{q}(\sigma_{\text{grp},g}^2)} \leftarrow \lambda_{\text{q}(\sigma_{\text{grp},g}^2)} + \left\| \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{grp},i}^g)} \right\|^2 + \text{tr} \left(\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{grp},i}^g)} \right) \\
& \lambda_{\text{q}(\sigma_{\text{gbl}}^2)} \leftarrow \mu_{\text{q}(1/a_{\text{gbl}})} + \left\| \boldsymbol{\mu}_{\text{q}(\mathbf{u}_{\text{gbl}})} \right\|^2 + \text{tr} \left(\boldsymbol{\Sigma}_{\text{q}(\mathbf{u}_{\text{gbl}})} \right) \\
& \mu_{\text{q}(1/\sigma_\varepsilon^2)} \leftarrow \xi_{\text{q}(\sigma_\varepsilon)} / \lambda_{\text{q}(\sigma_\varepsilon^2)} \quad ; \quad \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)} \leftarrow \xi_{\text{q}(\sigma_{\text{gbl}}^2)} / \lambda_{\text{q}(\sigma_{\text{gbl}}^2)} \\
& \mathbf{M}_{\text{q}((\boldsymbol{\Sigma}_g)^{-1})} \leftarrow (\xi_{\text{q}(\boldsymbol{\Sigma}_g)} - 2 + 1) \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma}_g)}^{-1} \quad ; \quad \mathbf{M}_{\text{q}((\boldsymbol{\Sigma}_h)^{-1})} \leftarrow (\xi_{\text{q}(\boldsymbol{\Sigma}_h)} - 2 + 1) \boldsymbol{\Lambda}_{\text{q}(\boldsymbol{\Sigma}_h)}^{-1} \\
& \mu_{\text{q}(1/\sigma_{\text{grp},g}^2)} \leftarrow \xi_{\text{q}(\sigma_{\text{grp},g}^2)} / \lambda_{\text{q}(\sigma_{\text{grp},g}^2)} \quad ; \quad \mu_{\text{q}(1/\sigma_{\text{grp},h}^2)} \leftarrow \xi_{\text{q}(\sigma_{\text{grp},h}^2)} / \lambda_{\text{q}(\sigma_{\text{grp},h}^2)} \\
& \lambda_{\text{q}(a_\varepsilon)} \leftarrow \mu_{\text{q}(1/\sigma_\varepsilon^2)} + 1/(\nu_\varepsilon s_\varepsilon^2) \quad ; \quad \mu_{\text{q}(1/a_\varepsilon)} \leftarrow \xi_{\text{q}(a_\varepsilon)} / \lambda_{\text{q}(a_\varepsilon)} \\
& \mathbf{M}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_g}^{-1})} \leftarrow \xi_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_g})} \boldsymbol{\Lambda}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_g})}^{-1} \quad ; \quad \mathbf{M}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_h}^{-1})} \leftarrow \xi_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_h})} \boldsymbol{\Lambda}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_h})}^{-1} \\
& \boldsymbol{\Lambda}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_g})} \leftarrow \text{diag}\{\text{diagonal}(\mathbf{M}_{\text{q}(\boldsymbol{\Sigma}_g^{-1})})\} + \{\nu_{\boldsymbol{\Sigma}_g} \text{diag}(s_{\boldsymbol{\Sigma}_g,1}^2, s_{\boldsymbol{\Sigma}_g,2}^2)\}^{-1} \\
& \boldsymbol{\Lambda}_{\text{q}(\mathbf{A}_{\boldsymbol{\Sigma}_h})} \leftarrow \text{diag}\{\text{diagonal}(\mathbf{M}_{\text{q}(\boldsymbol{\Sigma}_h^{-1})})\} + \{\nu_{\boldsymbol{\Sigma}_h} \text{diag}(s_{\boldsymbol{\Sigma}_h,1}^2, s_{\boldsymbol{\Sigma}_h,2}^2)\}^{-1} \\
& \lambda_{\text{q}(a_{\text{gbl}})} \leftarrow \mu_{\text{q}(1/\sigma_{\text{gbl}}^2)} + 1/(\nu_{\text{gbl}} s_{\text{gbl}}^2) \quad ; \quad \mu_{\text{q}(1/a_{\text{gbl}})} \leftarrow \xi_{\text{q}(a_{\text{gbl}})} / \lambda_{\text{q}(a_{\text{gbl}})} \\
& \lambda_{\text{q}(a_{\text{grp},g})} \leftarrow \mu_{\text{q}(1/\sigma_{\text{grp},g}^2)} + 1/(\nu_{\text{grp},g} s_{\text{grp},g}^2) \quad ; \quad \mu_{\text{q}(1/a_{\text{grp},g})} \leftarrow \xi_{\text{q}(a_{\text{grp},g})} / \lambda_{\text{q}(a_{\text{grp},g})} \\
& \lambda_{\text{q}(a_{\text{grp},h})} \leftarrow \mu_{\text{q}(1/\sigma_{\text{grp},h}^2)} + 1/(\nu_{\text{grp},h} s_{\text{grp},h}^2) \quad ; \quad \mu_{\text{q}(1/a_{\text{grp},h})} \leftarrow \xi_{\text{q}(a_{\text{grp},h})} / \lambda_{\text{q}(a_{\text{grp},h})}
\end{aligned}$$

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

continued on a subsequent page ...

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

$$\begin{aligned}
& \text{Outputs: } \boldsymbol{\mu}_{\mathbf{q}(\boldsymbol{\beta}, \mathbf{u}_{\text{gbl}})}, \boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\beta}, \mathbf{u}_{\text{gbl}})}, \left\{ \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)}, \boldsymbol{\Sigma}_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)}, \right. \\
& E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\boldsymbol{\beta}, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right)^T \right\} : 1 \leq i \leq m, \\
& E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\boldsymbol{\beta}, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\}, \\
& E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\}, \\
& \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)}, \boldsymbol{\Sigma}_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} : 1 \leq i \leq m, 1 \leq j \leq n_i \}, \xi_{\mathbf{q}(\sigma_{\varepsilon})}, \lambda_{\mathbf{q}(\sigma_{\varepsilon}^2)}, \xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)}, \\
& \lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}, \xi_{\mathbf{q}(\boldsymbol{\Sigma}_g)}, \boldsymbol{\Lambda}_{\mathbf{q}(\boldsymbol{\Sigma}_g)}^{-1}, \xi_{\mathbf{q}(\boldsymbol{\Sigma}_h)}, \boldsymbol{\Lambda}_{\mathbf{q}(\boldsymbol{\Sigma}_h)}^{-1}, \xi_{\mathbf{q}(\sigma_{\text{grp},g}^2)}, \lambda_{\mathbf{q}(\sigma_{\text{grp},g}^2)}, \xi_{\mathbf{q}(\sigma_{\text{grp},h}^2)}, \lambda_{\mathbf{q}(\sigma_{\text{grp},h}^2)}.
\end{aligned}$$

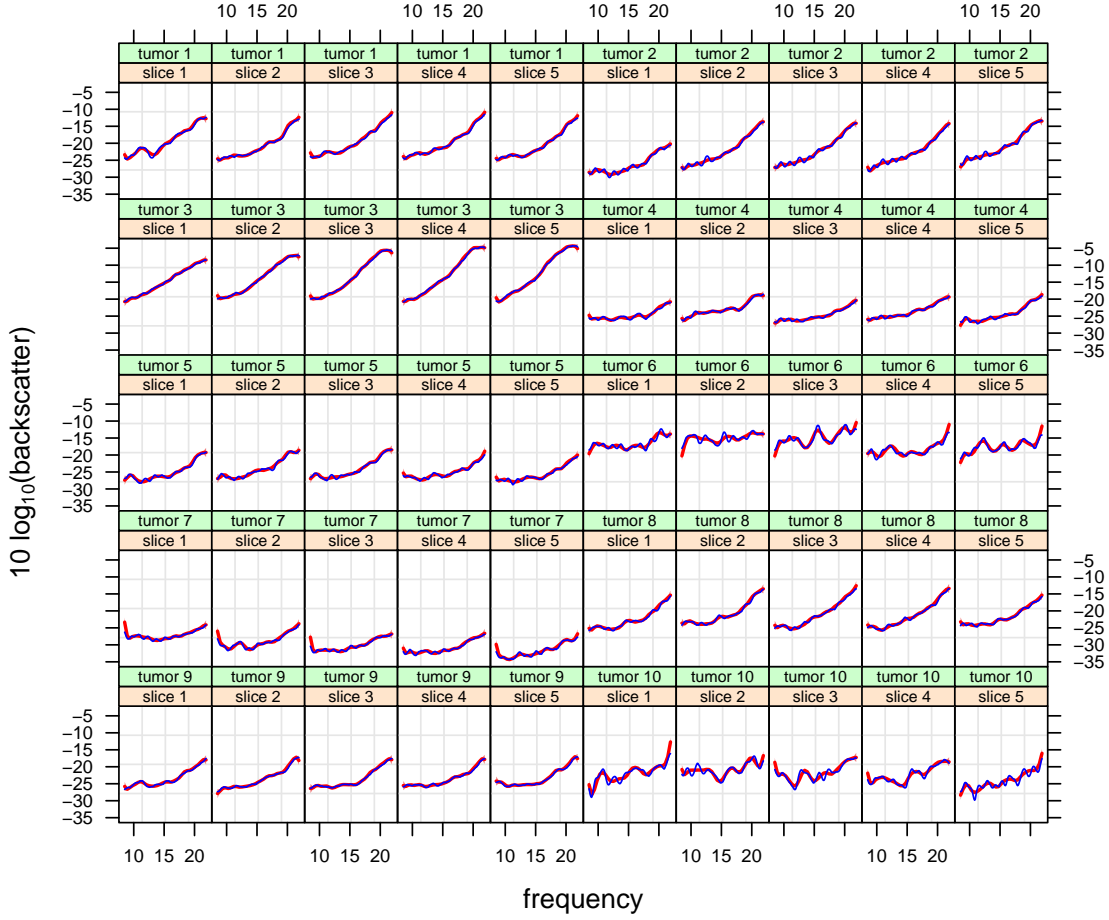


Figure 3: Illustrative three-level curve-type data with approximate fitted group specific curves and corresponding 99% credible sets based on mean field variational Bayes via Algorithm 4. The response variable is $10 \log_{10}(\text{backscatter})$ according to ultrasound technology. Level 1 corresponds to different ultrasound frequencies and matches the horizontal axes in each panel. Level 2 corresponds to different slices of a tumor due to differing probe locations. Level 3 corresponds to different tumors with one tumor for each of 10 laboratory mice.

4 Accuracy and Speed Assessment

In this section we provide some assessment of the accuracy and speed of the inference delivered by streamlined variational inference for group-specific curves models.

4.1 Accuracy Assessment

Mean field restrictions such as (11) and (20) imply that there is some loss of accuracy in inference produced by Algorithms 2 and 4. However, at least for the Gaussian response case treated here, approximate parameter orthogonality between the coefficient parameters and covariance parameters from likelihood theory implies that such restrictions are mild and mean field accuracy is high. Figure 4 corroborates this claim by assessing accuracy of the mean function estimates and 95% credible intervals at the median values of frequency for each panel in Figure 3. As a benchmark we use Markov chain Monte Carlo-based inference via the `rstan` package (Guo *et al.*, 2018). After a warmup of size 1,000 we retained 5,000 Markov chain Monte Carlo samples from the mean function and median frequency posterior distributions and used kernel density estimation to approximate the corresponding posterior density function. For a generic univariate parameter θ , the

accuracy of an approximation $q(\theta)$ to $p(\theta|\mathbf{y})$ is defined to be

$$\text{accuracy} \equiv 100 \left\{ 1 - \frac{1}{2} \int_{-\infty}^{\infty} |q(\theta) - p(\theta|\mathbf{y})| d\theta \right\} \%. \quad (23)$$

The percentages in the top right-hand panel of Figure 4 correspond to (23) with replacement of $p(\theta|\mathbf{y})$ by the above-mentioned kernel density estimate. In this case accuracy is seen to be excellent, with accuracy percentages between 97% and 99% for all 40 curves.

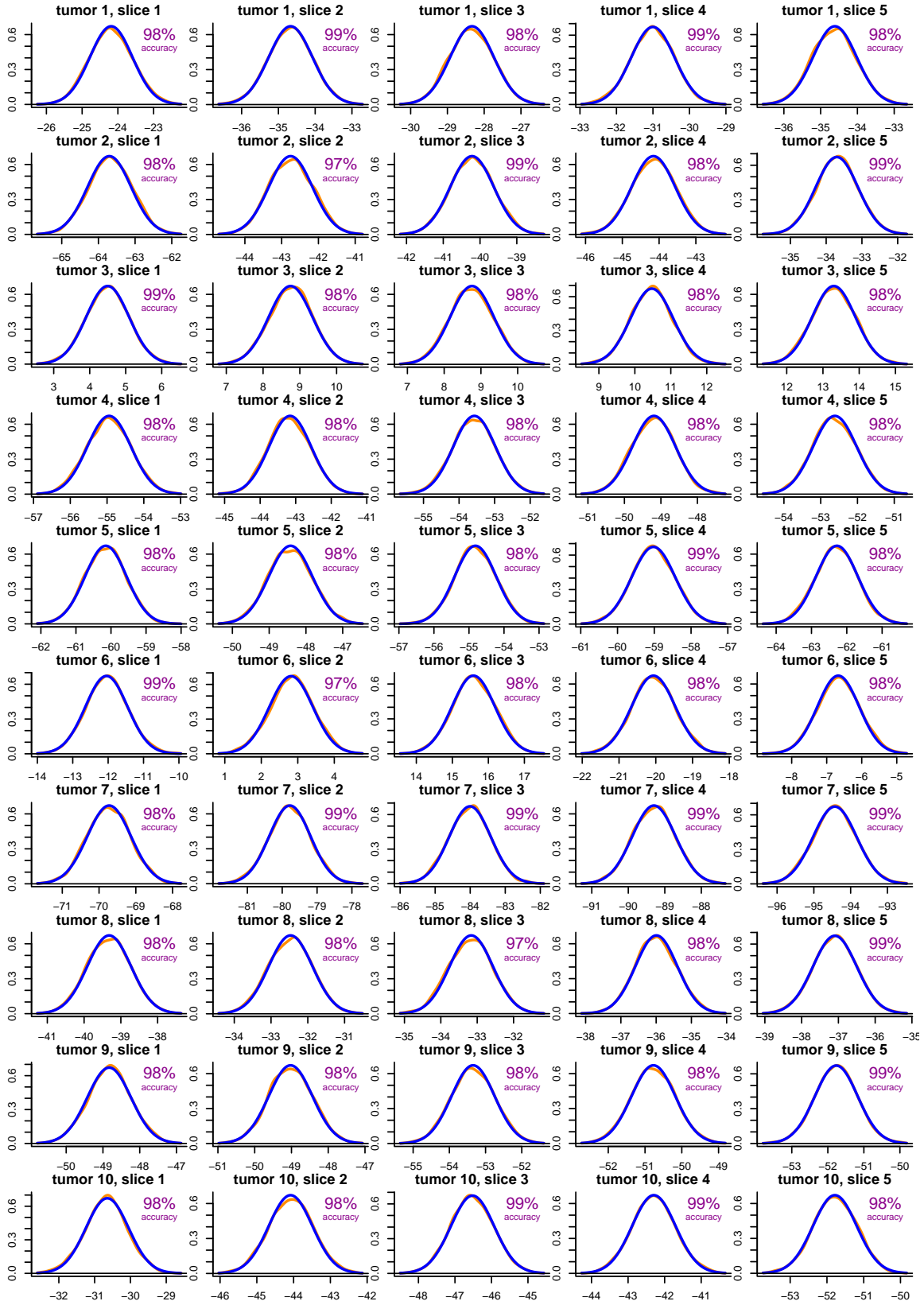


Figure 4: Accuracy assessment of Algorithm 4. Each panel displays approximate posterior density functions corresponding to mean function estimates according to the three-level group specific curve model (16). In each case the estimate is at the median frequency value. The orange density functions are based on Markov chain Monte Carlo and the blue density functions are based on mean field variational Bayes. The accuracy percentage scores are defined by (23).

4.2 Speed Assessment

We also conducted some simulation studies to assess the speed of streamlined variational higher level group-specific curve models, in terms of both comparative advantage over naïve implementation and absolute performance. The focus of these studies was variational inference in the two-level case and to probe maximal speed potential Algorithm 2 was implemented in the low-level computer language `Fortran` 77. An implementation of the naïve counterpart of Algorithm 2, involving storage and direct calculations concerning the full $\Sigma_{q(\beta, u)}$ matrix, was also carried out. We then simulated data according to model (1) with $\sigma_\varepsilon = 0.2$,

$$f(x) = 3\sqrt{x(1.3 - x)}\Phi(6x - 3) \quad \text{and} \quad g_i(x) = \alpha_1\alpha_2\sin(2\pi x^{\alpha_3})$$

where, for each i , α_1 , α_2 and α_3 are, respectively, random draws from the $N(\frac{1}{4}, \frac{1}{4})$ distribution and the sets $\{-1, 1\}$ and $\{1, 2, 3\}$. The level-2 sample sizes n_i generated randomly from the set $\{30, 31, \dots, 60\}$ and the level-1 sample sizes m ranging over the set $\{100, 200, 300, 400, 500\}$. All x_{ij} data were generated from a Uniform distribution over the unit interval. Table 1 summarizes the timings based on 100 replications with the number of mean field variational Bayes iterations fixed at 50. The study was run on a MacBook Air laptop with a 2.2 gigahertz processor and 8 gigabytes of random access memory.

m	naïve	streamlined	naïve/streamlined
100	75 (1.21)	0.748 (0.0334)	100
200	660 (7.72)	1.490 (0.0491)	442
300	2210 (22.00)	2.260 (0.0567)	974
400	5180 (92.20)	3.040 (0.0718)	1700
500	NA	3.780 (0.0593)	NA

Table 1: Average (standard deviation) of elapsed computing times in seconds for fitting model (1) naïvely versus with streamlining via Algorithm 2. The NA entries indicates non applicability due to the naïve computations not being feasible.

For m ranging from 100 to 400 we see that the naïve to streamlined ratios increase from about 100 to 1,700. When $m = 500$ the naïve implementation fails to run due to its excessive storage demands. In contrast, the streamlined fits are produced in about 3 seconds. It is clear that streamlined variational inference is to be preferred and is the only option for large numbers of groups.

We then obtained timings for the streamlined algorithm for m becoming much larger, taking on values 100, 500, 2,500 and 12,500. The iterations in Algorithm 2 were stopped when the relative increase in the marginal log-likelihood fell below 10^{-5} . The average and standard deviation times in seconds over 100 replications are shown in Table 2. We see that the computational times are approximately linear in m . Even with twelve and a half thousand groups, Algorithm 2 is able to deliver fitting and inference on a contemporary laptop computer in about one and a half minutes.

$m = 100$	$m = 500$	$m = 2,500$	$m = 12,500$
0.635	2.900	16.90	95.00
(0.183)	(0.391)	(1.92)	(4.92)

Table 2: Average (standard deviation) of elapsed computing times in seconds for fitting model (1) with streamlining via Algorithm 2.

Acknowledgments

This research was partially supported by the Australian Research Council Discovery Project DP140100441. The ultrasound data was provided by the Bioacoustics Research Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Illinois, U.S.A.

References

- Atay-Kayis, A. & Massam, H. (2005). A Monte Carlo method for computing marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, **92**, 317–335.
- Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015). Fitting linear mixed-effects models using `lme4`. *Journal of Statistical Software*, **67**(1), 1–48.
- Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.
- Brumback, B.A. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961–994.
- Donnelly, C.A., Laird, N.M. and Ware, J.H. (1995). Prediction and creation of smooth curves for temporally correlated longitudinal data. *Journal of the American Statistical Association*, **90**, 984–989.
- Durban, M., Harezlak, J., Wand, M.P. and Carroll, R.J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, **24**, 1153–1167.
- Goldsmith, J., Zipunnikov, V. and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, **71**, 344–353.
- Guo, J., Gabry, J. and Goodrich, B. (2018). `rstan`: R interface to Stan. R package version 2.18.2.
<http://mc-stan.org>.
- Huang, A. and Wand, M.P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, **8**, 439–452.
- Lee, C.Y.Y. and Wand, M.P. (2016). Variational inference for fitting complex Bayesian mixed effects models to health data. *Statistics in Medicine*, **35**, 165–188.
- Nolan, T.H., Menictas, M. and Wand, M.P. (2019). Streamlined computing for variational inference with higher level random effects. Unpublished manuscript submitted to the arXiv.org e-Print archive; on hold as of 11th March 2019. Soon to be posted also on <http://matt-wand.utsacademics.info/statsPapers.html>
- Nolan, T.H. and Wand, M.P. (2018). Solutions to sparse multilevel matrix problems. Unpublished manuscript available at <https://arxiv.org/abs/1903.03089>.
- Pinheiro, J.C. and Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.
- Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team. (2018). `nlme`: Linear and

nonlinear mixed effects models. R package version 3.1.
<http://cran.r-project.org/package=nlme>.

- Pratt, J.H., Jones, J.J., Miller, J.Z., Wagner, M.A. and Fineberg, N.S. (1989). Racial differences in aldosterone excretion and plasma aldosterone concentrations in children. *New England Journal of Medicine*, **321**, 1152–1157.
- Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–51.
- Trail, J.B., Collins, L.M., Rivera, D.E., Li, R., Piper, M.E. and Baker, T.B. (2014). Functional data analysis for dynamical system identification of behavioral processes. *Psychological Methods*, **19**(2), 175–187.
- Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics*, **48**, 269–312.
- Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.
- Wand, M.P. and Ormerod, J.T. (2008). On semiparametric regression with O’Sullivan penalized splines. *Australian and New Zealand Journal of Statistics*, **50**, 179–198.
- Wand, M.P. and Ormerod, J.T. (2011). Penalized wavelets: embedding wavelets into semi-parametric regression. *Electronic Journal of Statistics*, **5**, 1654–1717.
- Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Series B*, **60**, 159–174.
- Wirtzfeld, L.A., Ghoshal, G., Rosado-Mendez, I.M., Nam, K., Park, Y., Pawlicki, A.D., Miller, R.J., Simpson, D.G., Zagzebski, J.A., Oelze, M.I., Hall, T.J. and O’Brien, W.D. (2015). Quantitative ultrasound comparison of MAT and 4T1 mammary tumors in mice and rates across multiple imaging systems. *Journal of Ultrasound Medicine*, **34**, 1373–1383.
- Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, **93**, 710–719.

Web-Supplement for:
**Streamlined Variational Inference for
Higher Level Group-Specific Curve Models**

BY M. MENICTAS¹, T.H. NOLAN¹, D.G. SIMPSON² AND M.P. WAND¹

University of Technology Sydney¹ and University of Illinois²

S.1 Derivation of Result 1

Straightforward algebra can be used to verify that

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

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

$$b_i \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} y_i \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad B_i \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} X_i & \sigma_\varepsilon^{-1} Z_{\text{gbl},i} \\ O & m^{-1/2} \sigma_{\text{gbl}}^{-1} I_{K_{\text{gbl}}} \\ O & O \\ O & O \end{bmatrix} \quad \text{and} \quad \dot{B}_i \equiv \begin{bmatrix} \sigma_\varepsilon^{-1} X_i & \sigma_\varepsilon^{-1} Z_{\text{grp},i} \\ O & O \\ \Sigma^{-1/2} & O \\ O & \sigma_{\text{grp}}^{-1} I_{K_{\text{grp}}} \end{bmatrix}.$$

Therefore, in view of (6) and (7),

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

S.2 Derivation of Algorithm 1

Algorithm 1 is simply a proceduralization of Result 1.

S.3 The Inverse G-Wishart and Inverse χ^2 Distributions

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

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

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

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

if and only if the density function of X satisfies

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

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

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

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

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

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

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

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

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

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

S.4 Derivation of Result 2

It is straightforward to verify that the $\mu_{q(\beta, u)}$ and $\Sigma_{q(\beta, u)}$ updates, given at (12), may be written as

$$\mu_{q(\beta, u)} \leftarrow (B^T B)^{-1} B^T \mathbf{b} \quad \text{and} \quad \Sigma_{q(\beta, u)} \leftarrow (B^T B)^{-1}$$

where B and \mathbf{b} have the forms (9) with

$$\mathbf{b}_i \equiv \begin{bmatrix} \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{y}_i \\ m^{-1/2} \Sigma_{\beta}^{-1/2} \mu_{\beta} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad B_i \equiv \begin{bmatrix} \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{X}_i & \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{Z}_{\text{gbl}, i} \\ m^{-1/2} \Sigma_{\beta}^{-1/2} & \mathbf{O} \\ \mathbf{O} & m^{-1/2} \mu_{q(1/\sigma_{\text{gbl}}^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

and

$$\dot{B}_i \equiv \begin{bmatrix} \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{X}_i & \mu_{q(1/\sigma_{\varepsilon}^2)}^{1/2} \mathbf{Z}_{\text{grp}, i} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ M_{q(\Sigma^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & \mu_{q(1/\sigma_{\text{grp}}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}} \end{bmatrix}.$$

Result 2 immediately follows from Theorem 2 of Nolan & Wand (2018).

S.5 Derivation of Algorithm 2

We provide expressions for the q -densities for mean field variational Bayesian inference for the parameters in (10), with product density restriction (11). Arguments analogous to those given in, for example, Appendix C of Wand & Ormerod (2011) lead to:

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

where

$$\Sigma_{\mathbf{q}(\beta, \mathbf{u})} = (\mathbf{C}^T \mathbf{R}_{\text{MFVB}}^{-1} \mathbf{C} + \mathbf{D}_{\text{MFVB}})^{-1} \quad \text{and} \quad \mu_{\mathbf{q}(\beta, \mathbf{u})} = \Sigma_{\mathbf{q}(\beta, \mathbf{u})} (\mathbf{C}^T \mathbf{R}_{\text{MFVB}}^{-1} \mathbf{y} + \mathbf{o}_{\text{MFVB}})$$

with \mathbf{R}_{MFVB} , \mathbf{D}_{MFVB} and \mathbf{o}_{MFVB} defined via (13),

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

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

$$\begin{aligned} \lambda_{\mathbf{q}(\sigma_\varepsilon^2)} &= \mu_{\mathbf{q}(1/a_\varepsilon)} + \sum_{i=1}^m E_{\mathbf{q}} \left\{ \left\| \mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right\|^2 \right\} \\ &= \mu_{\mathbf{q}(1/a_\varepsilon)} + \sum_{i=1}^m \left[\left\| E_{\mathbf{q}} \left(\mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) \right\|^2 \right. \\ &\quad \left. + \text{tr} \left\{ \text{Cov}_{\mathbf{q}} \left(\mathbf{C}_{\text{gbl},i} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} + \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) \right\} \right] \\ &= \mu_{\mathbf{q}(1/a_\varepsilon)} + \sum_{i=1}^m \left\{ \left\| E_{\mathbf{q}} \left(\mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) \right\|^2 \right. \\ &\quad \left. + \text{tr}(\mathbf{C}_{\text{gbl},i}^T \mathbf{C}_{\text{gbl},i} \Sigma_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})}) + \text{tr}(\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{grp},i} \Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}) \right. \\ &\quad \left. + 2 \text{tr} \left[\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{gbl},i} E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \times \right. \right. \right. \\ &\quad \left. \left. \left. \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right)^T \right\} \right] \right\} \end{aligned}$$

where $\mathbf{C}_{\text{gbl},i} \equiv [\mathbf{X}_i \ \mathbf{Z}_{\text{gbl},i}]$, $\mathbf{C}_{\text{grp},i} \equiv [\mathbf{X}_i \ \mathbf{Z}_{\text{grp},i}]$, and with reciprocal moment $\mu_{\mathbf{q}(1/\sigma_\varepsilon^2)} = \xi_{\mathbf{q}(\sigma_\varepsilon^2)} / \lambda_{\mathbf{q}(\sigma_\varepsilon^2)}$,

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

where $\xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} = \nu_{\text{gbl}} + K_{\text{gbl}}$ and

$$\lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)} = \mu_{\mathbf{q}(1/a_{\text{gbl}})} + \|\mu_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}\|^2 + \text{tr}(\Sigma_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}),$$

with reciprocal moment $\mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} = \xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} / \lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}$,

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

where $\xi_{\mathbf{q}(\sigma_{\text{grp}}^2)} = \nu_{\text{grp}} + mK_{\text{grp}}$ and

$$\lambda_{\mathbf{q}(\sigma_{\text{grp}}^2)} = \mu_{\mathbf{q}(1/a_{\text{grp}})} + \sum_{i=1}^m \left\{ \|\mu_{\mathbf{q}(\mathbf{u}_{\text{grp},i})}\|^2 + \text{tr}(\Sigma_{\mathbf{q}(\mathbf{u}_{\text{grp},i})}) \right\},$$

with reciprocal moment $\mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} = \xi_{\mathbf{q}(\sigma_{\text{grp}}^2)} / \lambda_{\mathbf{q}(\sigma_{\text{grp}}^2)}$,

$$q(\Sigma) \text{ is an Inverse-G-Wishart}(G_{\text{full}}, \xi_{\mathbf{q}(\Sigma)}, \Lambda_{\mathbf{q}(\Sigma)}) \text{ density function}$$

where $\xi_{\mathbf{q}(\Sigma)} = \nu_{\Sigma} + 2 + m$

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

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

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

where $\xi_{\mathbf{q}(a_{\varepsilon})} = \nu_{\varepsilon} + 1$,

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

with reciprocal moment $\mu_{\mathbf{q}(1/a_{\varepsilon})} = \xi_{\mathbf{q}(a_{\varepsilon})}/\lambda_{\mathbf{q}(a_{\varepsilon})}$,

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

where $\xi_{\mathbf{q}(a_{\text{gbl}})} = \nu_{\text{gbl}} + 1$,

$$\lambda_{\mathbf{q}(a_{\text{gbl}})} = \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} + 1/(\nu_{\text{gbl}} s_{\text{gbl}}^2)$$

with reciprocal moment $\mu_{\mathbf{q}(1/a_{\text{gbl}})} = \xi_{\mathbf{q}(a_{\text{gbl}})}/\lambda_{\mathbf{q}(a_{\text{gbl}})}$,

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

where $\xi_{\mathbf{q}(a_{\text{grp}})} = \nu_{\text{grp}} + 1$,

$$\lambda_{\mathbf{q}(a_{\text{grp}})} = \mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} + 1/(\nu_{\text{grp}} s_{\text{grp}}^2)$$

with reciprocal moment $\mu_{\mathbf{q}(1/a_{\text{grp}})} = \xi_{\mathbf{q}(a_{\text{grp}})}/\lambda_{\mathbf{q}(a_{\text{grp}})}$ and

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

where $\xi_{\mathbf{q}(\mathbf{A}_{\Sigma})} = \nu_{\Sigma} + 2$,

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

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

S.6 Marginal Log-Likelihood Lower Bound and Derivation

The expression for the lower bound on the marginal log-likelihood for Algorithm 2 is

$$\begin{aligned}
\log \underline{\mathbf{p}}(\mathbf{y}; \mathbf{q}) = & \\
& -\frac{1}{2} \log(\pi) \sum_{i=1}^m n_i - \frac{1}{2} \log |\Sigma_{\beta}| - \frac{1}{2} \text{tr} \left(\Sigma_{\beta}^{-1} \left\{ \left(\mu_{\mathbf{q}(\beta)} - \mu_{\beta} \right) \left(\mu_{\mathbf{q}(\beta)} - \mu_{\beta} \right)^T + \Sigma_{\mathbf{q}(\beta)} \right\} \right) \\
& - \frac{1}{2} \text{tr} \left(\mathbf{M}_{\mathbf{q}(\Sigma^{-1})} \left\{ \sum_{i=1}^m \left(\mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i})} \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i})}^T + \Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},i})} \right) \right\} \right) + \frac{1}{2} \{2 + K_{\text{gbl}} + m(2 + K_{\text{grp}})\} \\
& - \frac{1}{2} \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} \left\{ \|\mu_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}\|^2 + \text{tr}(\Sigma_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}) \right\} - \frac{1}{2} \mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} \sum_{i=1}^m \left\{ \|\mu_{\mathbf{q}(\mathbf{u}_{\text{grp},i})}\|^2 + \text{tr}(\Sigma_{\mathbf{q}(\mathbf{u}_{\text{grp},i})}) \right\} \\
& + \frac{1}{2} \log |\Sigma_{\beta}| + \{\nu_{\Sigma} + m + 1 + \frac{1}{2}(\nu_{\varepsilon} + \nu_{\text{gbl}} + K_{\text{gbl}} + \nu_{\text{grp}} + mK_{\text{grp}})\} \log(2) - \log \Gamma(\frac{\nu_{\varepsilon}}{2}) \\
& - \frac{1}{2} \mu_{\mathbf{q}(1/a_{\varepsilon})} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)} - \frac{1}{2} \xi_{\mathbf{q}(\sigma_{\varepsilon}^2)} \log(\lambda_{\mathbf{q}(\sigma_{\varepsilon}^2)}) + \log \{\Gamma(\frac{1}{2} \xi_{\mathbf{q}(\sigma_{\varepsilon}^2)})\} + \frac{1}{2} \lambda_{\mathbf{q}(\sigma_{\varepsilon}^2)} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)} - \frac{1}{2} \log(\nu_{\varepsilon} s_{\varepsilon}^2) \\
& - 3 \log \{\Gamma(\frac{1}{2})\} - \frac{1}{2\nu_{\varepsilon} s_{\varepsilon}^2} \mu_{\mathbf{q}(1/a_{\varepsilon})} - \frac{1}{2} \xi_{\mathbf{q}(a_{\varepsilon})} \log(\lambda_{\mathbf{q}(a_{\varepsilon})}) + \log \{\Gamma(\frac{1}{2} \xi_{\mathbf{q}(a_{\varepsilon})})\} + \frac{1}{2} \lambda_{\mathbf{q}(a_{\varepsilon})} \mu_{\mathbf{q}(1/a_{\varepsilon})} \\
& - \log \Gamma(\frac{\nu_{\text{gbl}}}{2}) - \frac{1}{2} \mu_{\mathbf{q}(1/a_{\text{gbl}})} \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} - \frac{1}{2} \xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} \log(\lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}) + \log \{\Gamma(\frac{1}{2} \xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)})\} - \frac{1}{2} \log(\nu_{\text{gbl}} s_{\text{gbl}}^2) \\
& + \frac{1}{2} \lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)} \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} - \{1/(2\nu_{\text{gbl}} s_{\text{gbl}}^2)\} \mu_{\mathbf{q}(1/a_{\text{gbl}})} - \frac{1}{2} \xi_{\mathbf{q}(a_{\text{gbl}})} \log(\lambda_{\mathbf{q}(a_{\text{gbl}})}) - \frac{1}{2} \mu_{\mathbf{q}(1/a_{\text{grp}})} \mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} \\
& + \log \{\Gamma(\frac{1}{2} \xi_{\mathbf{q}(a_{\text{gbl}})})\} + \frac{1}{2} \lambda_{\mathbf{q}(a_{\text{gbl}})} \mu_{\mathbf{q}(1/a_{\text{gbl}})} - \log \Gamma(\frac{\nu_{\text{grp}}}{2}) + \log \{\Gamma(\frac{1}{2} \xi_{\mathbf{q}(\sigma_{\text{grp}}^2)})\} - \frac{1}{2} \log(\nu_{\text{grp}} s_{\text{grp}}^2) \\
& - \frac{1}{2} \xi_{\mathbf{q}(\sigma_{\text{grp}}^2)} \log(\lambda_{\mathbf{q}(\sigma_{\text{grp}}^2)}) + \frac{1}{2} \lambda_{\mathbf{q}(\sigma_{\text{grp}}^2)} \mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} - \{1/(2\nu_{\text{grp}} s_{\text{grp}}^2)\} \mu_{\mathbf{q}(1/a_{\text{grp}})} - \frac{1}{2} \xi_{\mathbf{q}(a_{\text{grp}})} \log(\lambda_{\mathbf{q}(a_{\text{grp}})}) \\
& + \log \{\Gamma(\frac{1}{2} \xi_{\mathbf{q}(a_{\text{grp}})})\} + \frac{1}{2} \lambda_{\mathbf{q}(a_{\text{grp}})} \mu_{\mathbf{q}(1/a_{\text{grp}})} - \frac{1}{2} \text{tr}(\mathbf{M}_{\mathbf{q}(\mathbf{A}_{\Sigma}^{-1})} \mathbf{M}_{\mathbf{q}(\Sigma^{-1})}) + \frac{1}{2} \text{tr}(\mathbf{A}_{\mathbf{q}(\Sigma)} \mathbf{M}_{\mathbf{q}(\Sigma^{-1})}) \\
& + \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(\xi_{\mathbf{q}(\mathbf{A}_{\Sigma})} + 2 - j)) - \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(\nu_{\Sigma} + 4 - j)) - \frac{1}{2}(\xi_{\mathbf{q}(\Sigma)} - 1) \log |\mathbf{A}_{\mathbf{q}(\Sigma)}| \\
& + \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(\xi_{\mathbf{q}(\Sigma)} + 2 - j)) - \frac{1}{2} \sum_{j=1}^2 1/(\nu_{\Sigma} s_{\Sigma,j}^2) \left(\mathbf{M}_{\mathbf{q}(\mathbf{A}_{\Sigma}^{-1})} \right)_{jj} - \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(3 - j)) \\
& - \frac{1}{2}(\xi_{\mathbf{q}(\mathbf{A}_{\Sigma})} - 1) \log |\mathbf{A}_{\mathbf{q}(\mathbf{A}_{\Sigma})}| + \frac{1}{2} \text{tr}(\mathbf{A}_{\mathbf{q}(\mathbf{A}_{\Sigma})} \mathbf{M}_{\mathbf{q}(\mathbf{A}_{\Sigma}^{-1})}) \\
& - \frac{1}{2} \mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)} \sum_{i=1}^m \left\{ \left\| E_{\mathbf{q}} \left(\mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl},i} \end{bmatrix} - \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) \right\|^2 \right. \\
& \quad + \text{tr}(\mathbf{C}_{\text{gbl},i}^T \mathbf{C}_{\text{gbl},i} \Sigma_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})}) + \text{tr}(\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{grp},i} \Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}) \\
& \quad \left. + 2 \text{tr} \left[\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{gbl},i} E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \times \right. \right. \right. \\
& \quad \left. \left. \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right)^T \right\} \right] \right\}. \tag{S.1}
\end{aligned}$$

Derivation: The lower-bound on the marginal log-likelihood is achieved through the following expression:

$$\begin{aligned}
\log \underline{\mathbf{p}}(\mathbf{y}; \mathbf{q}) = & E_{\mathbf{q}} \{ \log \mathbf{p}(\mathbf{y}, \beta, \mathbf{u}, \sigma_{\varepsilon}^2, a_{\varepsilon}, \sigma_{\text{gbl}}^2, a_{\text{gbl}}, \sigma_{\text{grp}}^2, a_{\text{grp}}, \Sigma, \mathbf{A}_{\Sigma}) \\
& - \log \mathbf{q}^*(\beta, \mathbf{u}, \sigma_{\varepsilon}^2, a_{\varepsilon}, \sigma_{\text{gbl}}^2, a_{\text{gbl}}, \sigma_{\text{grp}}^2, a_{\text{grp}}, \Sigma, \mathbf{A}_{\Sigma}) \}
\end{aligned}$$

$$\begin{aligned}
&= E_q\{\log \mathbf{p}(\mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2)\} \\
&\quad + E_q\{\log \mathbf{p}(\boldsymbol{\beta}, \mathbf{u} | \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \boldsymbol{\Sigma})\} - E_q\{\log \mathbf{q}^*(\boldsymbol{\beta}, \mathbf{u})\} \\
&\quad + E_q\{\log \mathbf{p}(\sigma_\varepsilon^2 | a_\varepsilon)\} - E_q\{\log \mathbf{q}^*(\sigma_\varepsilon^2)\} + E_q\{\log \mathbf{p}(a_\varepsilon)\} - E_q\{\log \mathbf{q}^*(a_\varepsilon)\} \\
&\quad + E_q\{\log \mathbf{p}(\sigma_{\text{gbl}}^2 | a_{\text{gbl}})\} - E_q\{\log \mathbf{q}^*(\sigma_{\text{gbl}}^2)\} + E_q\{\log \mathbf{p}(a_{\text{gbl}})\} - E_q\{\log \mathbf{q}^*(a_{\text{gbl}})\} \\
&\quad + E_q\{\log \mathbf{p}(\sigma_{\text{grp}}^2 | a_{\text{grp}})\} - E_q\{\log \mathbf{q}^*(\sigma_{\text{grp}}^2)\} + E_q\{\log \mathbf{p}(a_{\text{grp}})\} - E_q\{\log \mathbf{q}^*(a_{\text{grp}})\} \\
&\quad + E_q\{\log \mathbf{p}(\boldsymbol{\Sigma} | \mathbf{A}_\Sigma)\} - E_q\{\log \mathbf{q}^*(\boldsymbol{\Sigma})\} + E_q\{\log \mathbf{p}(\mathbf{A}_\Sigma)\} - E_q\{\log \mathbf{q}^*(\mathbf{A}_\Sigma)\}.
\end{aligned} \tag{S.2}$$

First we note that

$$\log \mathbf{p}(\mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2) = -\frac{1}{2} \log(2\pi) \sum_{i=1}^m n_i - \frac{1}{2} \log(\sigma_\varepsilon^2) \sum_{i=1}^m n_i - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^m \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2$$

where

$$\begin{aligned}
&\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 \\
&= \left\| \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} - \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} \boldsymbol{\beta} - \begin{bmatrix} \mathbf{Z}_{\text{gbl},1} \\ \vdots \\ \mathbf{Z}_{\text{gbl},m} \end{bmatrix} \mathbf{u}_{\text{gbl}} - \underset{1 \leq i \leq m}{\text{blockdiag}}([\mathbf{X}_i \ \mathbf{Z}_{\text{grp},i}]) \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix}_{1 \leq i \leq m} \right\|^2 \\
&= \sum_{i=1}^m \|\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_{\text{gbl},i} \mathbf{u}_{\text{gbl},i} - \mathbf{X}_i \mathbf{u}_{\text{lin},i} - \mathbf{Z}_{\text{grp},i} \mathbf{u}_{\text{grp},i}\|^2 \\
&= \sum_{i=1}^m \left\| \mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}_{\text{gbl},i} \end{bmatrix} - \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right\|^2
\end{aligned}$$

and

$$\mathbf{C}_{\text{gbl},i} \equiv [\mathbf{X}_i \ \mathbf{Z}_{\text{gbl},i}], \quad \mathbf{C}_{\text{grp},i} \equiv [\mathbf{X}_i \ \mathbf{Z}_{\text{grp},i}].$$

Therefore,

$$\begin{aligned}
&E_q\{\log \mathbf{p}(\mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2)\} \\
&= -\frac{1}{2} \log(2\pi) \sum_{i=1}^m n_i - \frac{1}{2} E_q\{\log(\sigma_\varepsilon^2)\} \sum_{i=1}^m n_i \\
&\quad - \frac{1}{2} \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)} \sum_{i=1}^m \left\{ \left\| E_q \left(\mathbf{y}_i - \mathbf{C}_{\text{gbl},i} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}_{\text{gbl},i} \end{bmatrix} - \mathbf{C}_{\text{grp},i} \begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} \right) \right\|^2 \right. \\
&\quad \left. + \text{tr}(\mathbf{C}_{\text{gbl},i}^T \mathbf{C}_{\text{gbl},i} \boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\beta}, \mathbf{u}_{\text{gbl}})}) + \text{tr}(\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{grp},i} \boldsymbol{\Sigma}_{\mathbf{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})}) \right. \\
&\quad \left. + 2 \text{tr} \left[\mathbf{C}_{\text{grp},i}^T \mathbf{C}_{\text{gbl},i} E_q \left\{ \left(\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\boldsymbol{\beta}, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i} \\ \mathbf{u}_{\text{grp},i} \end{bmatrix} - \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},i}, \mathbf{u}_{\text{grp},i})} \right)^T \right\} \right] \right\}
\end{aligned}$$

The remainder of the expectations in (S.2) are expressed as:

$$\begin{aligned}
E_q\{\log \mathbf{p}(\boldsymbol{\beta}, \mathbf{u} | \sigma_{\text{gbl}}^2, \sigma_{\text{grp}}^2, \boldsymbol{\Sigma})\} &= -\frac{1}{2} \{2 + K_{\text{gbl}} + m(2 + K_{\text{grp}})\} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}_\beta| \\
&\quad - \frac{K_{\text{gbl}}}{2} E_q\{\log(\sigma_{\text{gbl}}^2)\} - \frac{m}{2} E_q\{\log |\boldsymbol{\Sigma}|\} - \frac{mK_{\text{grp}}}{2} E_q\{\log(\sigma_{\text{grp}}^2)\} \\
&\quad - \frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}_\beta^{-1} \left\{ \left(\boldsymbol{\mu}_{\mathbf{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_\beta \right) \left(\boldsymbol{\mu}_{\mathbf{q}(\boldsymbol{\beta})} - \boldsymbol{\mu}_\beta \right)^T + \boldsymbol{\Sigma}_{\mathbf{q}(\boldsymbol{\beta})} \right\} \right) \\
&\quad - \frac{1}{2} \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} \left\{ \|\boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}) \right\} \\
&\quad - \frac{1}{2} \text{tr} \left(\mathbf{M}_{\mathbf{q}(\boldsymbol{\Sigma}^{-1})} \left\{ \sum_{i=1}^m \left(\boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},i})} \boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{lin},i})}^T + \boldsymbol{\Sigma}_{\mathbf{q}(\mathbf{u}_{\text{lin},i})} \right) \right\} \right) \\
&\quad - \frac{1}{2} \mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} \sum_{i=1}^m \left\{ \|\boldsymbol{\mu}_{\mathbf{q}(\mathbf{u}_{\text{grp},i})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\mathbf{q}(\mathbf{u}_{\text{grp},i})}) \right\}
\end{aligned}$$

$$\begin{aligned}
E_q\{\log \mathbf{q}^*(\boldsymbol{\beta}, \mathbf{u})\} &= -\frac{1}{2}\{2 + K_{\text{gbl}} + m(2 + K_{\text{grp}})\} - \frac{1}{2}\{2 + K_{\text{gbl}} + m(2 + K_{\text{grp}})\} \log(2\pi) \\
&\quad - \frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \\
E_q\{\log \mathbf{p}(\sigma_{\varepsilon}^2 | a_{\varepsilon})\} &= -\frac{1}{2}\nu_{\varepsilon} E_q\{\log(2a_{\varepsilon})\} - \log \Gamma(\nu_{\varepsilon}/2) - (\frac{1}{2}\nu_{\varepsilon} + 1)E_q\{\log(\sigma_{\varepsilon}^2)\} \\
&\quad - \frac{1}{2}\mu_{\mathbf{q}(1/a_{\varepsilon})}\mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)} \\
E_q\{\log \mathbf{q}^*(\sigma_{\varepsilon}^2)\} &= \frac{1}{2}\xi_{\mathbf{q}(\sigma_{\varepsilon}^2)} \log(\lambda_{\mathbf{q}(\sigma_{\varepsilon}^2)}/2) - \log\{\Gamma(\frac{1}{2}\xi_{\mathbf{q}(\sigma_{\varepsilon}^2)})\} - (\frac{1}{2}\xi_{\mathbf{q}(\sigma_{\varepsilon}^2)} + 1)E_q\{\log(\sigma_{\varepsilon}^2)\} \\
&\quad - \frac{1}{2}\lambda_{\mathbf{q}(\sigma_{\varepsilon}^2)}\mu_{\mathbf{q}(1/\sigma_{\varepsilon}^2)} \\
E_q\{\log \mathbf{p}(a_{\varepsilon})\} &= -\frac{1}{2} \log(2\nu_{\varepsilon}s_{\varepsilon}^2) - \log\{\Gamma(\frac{1}{2})\} - (\frac{1}{2} + 1)E_q\{\log(a_{\varepsilon})\} \\
&\quad - \{1/(2\nu_{\varepsilon}s_{\varepsilon}^2)\}\mu_{\mathbf{q}(1/a_{\varepsilon})} \\
E_q\{\log \mathbf{q}^*(a_{\varepsilon})\} &= \frac{1}{2}\xi_{\mathbf{q}(a_{\varepsilon})} \log(\lambda_{\mathbf{q}(a_{\varepsilon})}/2) - \log\{\Gamma(\frac{1}{2}\xi_{\mathbf{q}(a_{\varepsilon})})\} - (\frac{1}{2}\xi_{\mathbf{q}(a_{\varepsilon})} + 1)E_q\{\log(a_{\varepsilon})\} \\
&\quad - \frac{1}{2}\lambda_{\mathbf{q}(a_{\varepsilon})}\mu_{\mathbf{q}(1/a_{\varepsilon})} \\
E_q\{\log \mathbf{p}(\sigma_{\text{gbl}}^2 | a_{\text{gbl}})\} &= -\frac{1}{2}\nu_{\text{gbl}} E_q\{\log(2a_{\text{gbl}})\} - \log \Gamma(\nu_{\text{gbl}}/2) - (\frac{1}{2}\nu_{\text{gbl}} + 1)E_q\{\log(\sigma_{\text{gbl}}^2)\} \\
&\quad - \frac{1}{2}\mu_{\mathbf{q}(1/a_{\text{gbl}})}\mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} \\
E_q\{\log \mathbf{q}^*(\sigma_{\text{gbl}}^2)\} &= \frac{1}{2}\xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} \log(\lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}/2) - \log\{\Gamma(\frac{1}{2}\xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)})\} - (\frac{1}{2}\xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} + 1)E_q\{\log(\sigma_{\text{gbl}}^2)\} \\
&\quad - \frac{1}{2}\lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}\mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} \\
E_q\{\log \mathbf{p}(a_{\text{gbl}})\} &= -\frac{1}{2} \log(2\nu_{\text{gbl}}s_{\text{gbl}}^2) - \log\{\Gamma(\frac{1}{2})\} - (\frac{1}{2} + 1)E_q\{\log(a_{\text{gbl}})\} \\
&\quad - \{1/(2\nu_{\text{gbl}}s_{\text{gbl}}^2)\}\mu_{\mathbf{q}(1/a_{\text{gbl}})} \\
E_q\{\log \mathbf{q}^*(a_{\text{gbl}})\} &= \frac{1}{2}\xi_{\mathbf{q}(a_{\text{gbl}})} \log(\lambda_{\mathbf{q}(a_{\text{gbl}})}/2) - \log\{\Gamma(\frac{1}{2}\xi_{\mathbf{q}(a_{\text{gbl}})})\} - (\frac{1}{2}\xi_{\mathbf{q}(a_{\text{gbl}})} + 1)E_q\{\log(a_{\text{gbl}})\} \\
&\quad - \frac{1}{2}\lambda_{\mathbf{q}(a_{\text{gbl}})}\mu_{\mathbf{q}(1/a_{\text{gbl}})} \\
E_q\{\log \mathbf{p}(\sigma_{\text{grp}}^2 | a_{\text{grp}})\} &= -\frac{1}{2}\nu_{\text{grp}} E_q\{\log(2a_{\text{grp}})\} - \log \Gamma(\nu_{\text{grp}}/2) - (\frac{1}{2}\nu_{\text{grp}} + 1)E_q\{\log(\sigma_{\text{grp}}^2)\} \\
&\quad - \frac{1}{2}\mu_{\mathbf{q}(1/a_{\text{grp}})}\mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} \\
E_q\{\log \mathbf{q}^*(\sigma_{\text{grp}}^2)\} &= \frac{1}{2}\xi_{\mathbf{q}(\sigma_{\text{grp}}^2)} \log(\lambda_{\mathbf{q}(\sigma_{\text{grp}}^2)}/2) - \log\{\Gamma(\frac{1}{2}\xi_{\mathbf{q}(\sigma_{\text{grp}}^2)})\} - (\frac{1}{2}\xi_{\mathbf{q}(\sigma_{\text{grp}}^2)} + 1)E_q\{\log(\sigma_{\text{grp}}^2)\} \\
&\quad - \frac{1}{2}\lambda_{\mathbf{q}(\sigma_{\text{grp}}^2)}\mu_{\mathbf{q}(1/\sigma_{\text{grp}}^2)} \\
E_q\{\log \mathbf{p}(a_{\text{grp}})\} &= -\frac{1}{2} \log(2\nu_{\text{grp}}s_{\text{grp}}^2) - \log\{\Gamma(\frac{1}{2})\} - (\frac{1}{2} + 1)E_q\{\log(a_{\text{grp}})\} \\
&\quad - \{1/(2\nu_{\text{grp}}s_{\text{grp}}^2)\}\mu_{\mathbf{q}(1/a_{\text{grp}})} \\
E_q\{\log \mathbf{q}^*(a_{\text{grp}})\} &= \frac{1}{2}\xi_{\mathbf{q}(a_{\text{grp}})} \log(\lambda_{\mathbf{q}(a_{\text{grp}})}/2) - \log\{\Gamma(\frac{1}{2}\xi_{\mathbf{q}(a_{\text{grp}})})\} - (\frac{1}{2}\xi_{\mathbf{q}(a_{\text{grp}})} + 1)E_q\{\log(a_{\text{grp}})\} \\
&\quad - \frac{1}{2}\lambda_{\mathbf{q}(a_{\text{grp}})}\mu_{\mathbf{q}(1/a_{\text{grp}})} \\
E_q[\log\{\mathbf{p}(\boldsymbol{\Sigma} | \mathbf{A}_{\boldsymbol{\Sigma}})\}] &= -\frac{1}{2}(\nu_{\boldsymbol{\Sigma}} + 1)E_q\{\log |\mathbf{A}_{\boldsymbol{\Sigma}}|\} - \frac{1}{2}(\nu_{\boldsymbol{\Sigma}} + 4)E_q\{\log |\boldsymbol{\Sigma}|\} - \frac{1}{2} \log(\pi) \\
&\quad - \frac{1}{2}\text{tr}(\mathbf{M}_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}}^{-1})}\mathbf{M}_{\mathbf{q}(\boldsymbol{\Sigma}^{-1})}) - (\nu_{\boldsymbol{\Sigma}} + 3) \log(2) - \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(\nu_{\boldsymbol{\Sigma}} + 4 - j)) \\
E_q[\log\{\mathbf{q}(\boldsymbol{\Sigma})\}] &= \frac{1}{2}(\xi_{\mathbf{q}(\boldsymbol{\Sigma})} - 1) \log |\boldsymbol{\Lambda}_{\mathbf{q}(\boldsymbol{\Sigma})}| - \frac{1}{2}(\xi_{\mathbf{q}(\boldsymbol{\Sigma})} + 2)E_q\{\log |\boldsymbol{\Sigma}|\} - \frac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathbf{q}(\boldsymbol{\Sigma})}\mathbf{M}_{\mathbf{q}(\boldsymbol{\Sigma}^{-1})}) \\
&\quad - (\xi_{\mathbf{q}(\boldsymbol{\Sigma})} + 1) \log(2) - \frac{1}{2} \log(\pi) - \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(\xi_{\mathbf{q}(\boldsymbol{\Sigma})} + 2 - j)) \\
E_q[\log\{\mathbf{p}(\mathbf{A}_{\boldsymbol{\Sigma}})\}] &= -\frac{3}{2}E_q\{\log |\mathbf{A}_{\boldsymbol{\Sigma}}|\} - \frac{1}{2} \sum_{j=1}^2 1/(\nu_{\boldsymbol{\Sigma}}s_{\boldsymbol{\Sigma},j}^2) \left(\mathbf{M}_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}}^{-1})}\right)_{jj} - 2 \log(2) - \frac{1}{2} \log(\pi) \\
&\quad - \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(3 - j)) \\
E_q[\log\{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})\}] &= \frac{1}{2}(\xi_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})} - 1) \log |\boldsymbol{\Lambda}_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})}| - \frac{1}{2}(\xi_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})} + 2)E_q\{\log |\mathbf{A}_{\boldsymbol{\Sigma}}|\} - \frac{1}{2}\text{tr}(\boldsymbol{\Lambda}_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})}\mathbf{M}_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}}^{-1})}) \\
&\quad - (\xi_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})} + 1) \log(2) - \frac{1}{2} \log(\pi) - \sum_{j=1}^2 \log \Gamma(\frac{1}{2}(\xi_{\mathbf{q}(\mathbf{A}_{\boldsymbol{\Sigma}})} + 2 - j))
\end{aligned}$$

In the summation of each of these $\log \underline{p}(\mathbf{y}; \mathbf{q})$ terms, note that the coefficient of $E_{\mathbf{q}}\{\log(\sigma_{\varepsilon}^2)\}$ is

$$-\frac{1}{2} \sum_{i=1}^m n_i - \frac{1}{2} \nu_{\varepsilon} - 1 + \frac{1}{2} \xi_{\mathbf{q}}(\sigma_{\varepsilon}^2) + 1 = -\frac{1}{2} \sum_{i=1}^m n_i - \frac{1}{2} \nu_{\varepsilon} - 1 + \frac{1}{2} (\nu_{\varepsilon} + \sum_{i=1}^m n_i) + 1 = 0.$$

The coefficient of $E_{\mathbf{q}}\{\log(\sigma_{\text{gbl}}^2)\}$ is

$$-\frac{1}{2} K_{\text{gbl}} - \frac{1}{2} \nu_{\text{gbl}} - 1 + \frac{1}{2} \xi_{\mathbf{q}}(\sigma_{\text{gbl}}^2) + 1 = -\frac{1}{2} K_{\text{gbl}} - \frac{1}{2} \nu_{\text{gbl}} - 1 + \frac{1}{2} (\nu_{\text{gbl}} + K_{\text{gbl}}) + 1 = 0.$$

The coefficient of $E_{\mathbf{q}}\{\log(\sigma_{\text{grp}}^2)\}$ is

$$-\frac{1}{2} m K_{\text{grp}} - \frac{1}{2} \nu_{\text{grp}} - 1 + \frac{1}{2} \xi_{\mathbf{q}}(\sigma_{\text{grp}}^2) + 1 = -\frac{1}{2} m K_{\text{grp}} - \frac{1}{2} \nu_{\text{grp}} - 1 + \frac{1}{2} (\nu_{\text{grp}} + m K_{\text{grp}}) + 1 = 0.$$

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

$$-\frac{m}{2} - \frac{1}{2} (\nu_{\Sigma} + 4) + \frac{1}{2} (\xi_{\mathbf{q}}(\Sigma) + 2) = -\frac{1}{2} (m + \nu_{\Sigma} + 4) + \frac{1}{2} (m + \nu_{\Sigma} + 4) = 0.$$

The coefficient of $E_{\mathbf{q}}\{\log(a_{\varepsilon})\}$ is

$$-\frac{1}{2} \nu_{\varepsilon} - \frac{1}{2} - 1 + \frac{1}{2} \xi_{\mathbf{q}}(a_{\varepsilon}) + 1 = -\frac{1}{2} \nu_{\varepsilon} - \frac{1}{2} - 1 + \frac{1}{2} (\nu_{\varepsilon} + 1) + 1 = 0.$$

The coefficient of $E_{\mathbf{q}}\{\log(a_{\text{gbl}})\}$ is

$$-\frac{1}{2} \nu_{\text{gbl}} - \frac{1}{2} - 1 + \frac{1}{2} \xi_{\mathbf{q}}(a_{\text{gbl}}) + 1 = -\frac{1}{2} \nu_{\text{gbl}} - \frac{1}{2} - 1 + \frac{1}{2} (\nu_{\text{gbl}} + 1) + 1 = 0.$$

The coefficient of $E_{\mathbf{q}}\{\log(a_{\text{grp}})\}$ is

$$-\frac{1}{2} \nu_{\text{grp}} - \frac{1}{2} - 1 + \frac{1}{2} \xi_{\mathbf{q}}(a_{\text{grp}}) + 1 = -\frac{1}{2} \nu_{\text{grp}} - \frac{1}{2} - 1 + \frac{1}{2} (\nu_{\text{grp}} + 1) + 1 = 0.$$

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

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

Therefore these terms can be dropped and then the cancellations led by the above expectations leads to the lower bound expression in (S.1).

S.7 Derivation of Result 3

If \mathbf{B} and \mathbf{b} have the same forms given by equation (7) in Nolan & Wand (2018) with

$$\mathbf{b}_{ij} \equiv \begin{bmatrix} \sigma_{\varepsilon}^{-1} \mathbf{y}_{ij} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \mathbf{B}_{ij} \equiv \begin{bmatrix} \sigma_{\varepsilon}^{-1} \mathbf{X}_{ij} & \sigma_{\varepsilon}^{-1} \mathbf{Z}_{\text{gbl},ij} \\ \mathbf{O} & (\sum_{i=1}^m n_i)^{-1/2} \sigma_{\text{gbl}}^{-1} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{\mathbf{B}}_{ij} \equiv \begin{bmatrix} \sigma_{\varepsilon}^{-1} \mathbf{X}_{ij} & \sigma_{\varepsilon}^{-1} \mathbf{Z}_{\text{grp},ij}^g \\ \mathbf{O} & \mathbf{O} \\ n_i^{-1/2} \Sigma_g^{-1/2} & \mathbf{O} \\ \mathbf{O} & n_i^{-1/2} \sigma_{\text{grp},g}^{-1} \mathbf{I}_{K_{\text{grp}}^g} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \text{and} \quad \ddot{\mathbf{B}}_{ij} \equiv \begin{bmatrix} \sigma_{\varepsilon}^{-1} \mathbf{X}_{ij} & \sigma_{\varepsilon}^{-1} \mathbf{Z}_{\text{grp},ij}^h \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \Sigma_h^{-1/2} & \mathbf{O} \\ \mathbf{O} & \sigma_{\text{grp},h}^{-1} \mathbf{I}_{K_{\text{grp}}^h} \end{bmatrix},$$

then straightforward algebra leads to

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

where

$$C \equiv [\mathbf{X} \ \mathbf{Z}], \quad D_{\text{BLUP}} \equiv \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{G}^{-1} \end{bmatrix} \quad \text{and} \quad R_{\text{BLUP}} \equiv \sigma_\varepsilon^2 \mathbf{I}, \quad (\text{S.3})$$

and \mathbf{G} as defined in (18). The remainder of the derivation of Result 3 is analogous to that of Result 1.

S.8 Derivation of Algorithm 3

Algorithm 3 is simply a proceduralization of Result 3.

S.9 Derivation of Result 4

It is straightforward to verify that the $\mu_{\mathbf{q}(\beta, \mathbf{u})}$ and $\Sigma_{\mathbf{q}(\beta, \mathbf{u})}$ updates, given at (12) but with D_{MFVB} as given in (21), may be written as

$$\mu_{\mathbf{q}(\beta, \mathbf{u})} \leftarrow (B^T B)^{-1} B^T \mathbf{b} \quad \text{and} \quad \Sigma_{\mathbf{q}(\beta, \mathbf{u})} \leftarrow (B^T B)^{-1}$$

where B and \mathbf{b} have the forms given by equation (7) in Nolan & Wand (2018) with

$$\mathbf{b}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{y}_{ij} \\ (\sum_{i=1}^m n_i)^{-1/2} \Sigma_\beta^{-1/2} \mu_\beta \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad B_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{gbl},ij} \\ (\sum_{i=1}^m n_i)^{-1/2} \Sigma_\beta^{-1/2} & \mathbf{O} \\ \mathbf{O} & (\sum_{i=1}^m n_i)^{-1/2} \mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)}^{1/2} \mathbf{I}_{K_{\text{gbl}}} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix},$$

$$\dot{B}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{grp},ij}^g \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ n_i^{-1/2} M_{\mathbf{q}(\Sigma_g^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & n_i^{-1/2} \mu_{\mathbf{q}(1/\sigma_{\text{grp},g}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}^g} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \text{and} \quad \ddot{B}_{ij} \equiv \begin{bmatrix} \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{X}_{ij} & \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{1/2} \mathbf{Z}_{\text{grp},ij}^h \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \\ M_{\mathbf{q}(\Sigma_h^{-1})}^{1/2} & \mathbf{O} \\ \mathbf{O} & \mu_{\mathbf{q}(1/\sigma_{\text{grp},h}^2)}^{1/2} \mathbf{I}_{K_{\text{grp}}^h} \end{bmatrix}.$$

Result 4 immediately follows from Theorem 4 of Nolan & Wand (2018).

S.10 Derivation of Algorithm 4

We provide expressions for the \mathbf{q} -densities for mean field variational Bayesian inference for the parameters in (19) with product density restriction (20).

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

where

$$\Sigma_{\mathbf{q}(\beta, \mathbf{u})} = (\mathbf{C}^T \mathbf{R}_{\text{MFVB}}^{-1} \mathbf{C} + \mathbf{D}_{\text{MFVB}})^{-1} \quad \text{and} \quad \mu_{\mathbf{q}(\beta, \mathbf{u})} = \Sigma_{\mathbf{q}(\beta, \mathbf{u})} (\mathbf{C}^T \mathbf{R}_{\text{MFVB}}^{-1} \mathbf{y} + \mathbf{o}_{\text{MFVB}})$$

$$\text{with } \mathbf{R}_{\text{MFVB}} \equiv \mu_{\mathbf{q}(1/\sigma_\varepsilon^2)}^{-1} \mathbf{I}, \mathbf{o}_{\text{MFVB}} \equiv \begin{bmatrix} \Sigma_\beta^{-1} \mu_\beta \\ \mathbf{0} \end{bmatrix} \text{ and } \mathbf{D}_{\text{MFVB}} \text{ as given in (21).}$$

$\mathbf{q}(\sigma_\varepsilon^2)$ is an Inverse- χ^2 ($\xi_{\mathbf{q}(\sigma_\varepsilon^2)}, \lambda_{\mathbf{q}(\sigma_\varepsilon^2)}$) density function

where $\xi_{\mathbf{q}(\sigma_\varepsilon^2)} = \nu_\varepsilon + \sum_{i=1}^m \sum_{j=1}^{n_i} o_{ij}$ and

$$\begin{aligned} \lambda_{\mathbf{q}(\sigma_\varepsilon^2)} &= \mu_{\mathbf{q}(1/a_\varepsilon)} + \sum_{i=1}^m \sum_{j=1}^{n_i} E_{\mathbf{q}} \left\{ \left\| \mathbf{y}_{ij} - \mathbf{C}_{\text{gbl},ij} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mathbf{C}_{\text{grp},ij}^g \begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mathbf{C}_{\text{grp},ij}^h \begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right\|^2 \right\} \\ &= \mu_{\mathbf{q}(1/a_\varepsilon)} + \sum_{i=1}^m \sum_{j=1}^{n_i} \left[\left\| E_{\mathbf{q}} \left(\mathbf{y}_{ij} - \mathbf{C}_{\text{gbl},ij} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mathbf{C}_{\text{grp},ij}^g \begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mathbf{C}_{\text{grp},ij}^h \begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right) \right\|^2 \right. \\ &\quad \left. + \text{tr} \left\{ \text{Cov}_{\mathbf{q}} \left(\mathbf{C}_{\text{gbl},ij} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} + \mathbf{C}_{\text{grp},ij}^g \begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} + \mathbf{C}_{\text{grp},ij}^h \begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right) \right\} \right] \\ &= \mu_{\mathbf{q}(1/a_\varepsilon)} + \sum_{i=1}^m \sum_{j=1}^{n_i} \left\{ \left\| E_{\mathbf{q}} \left(\mathbf{y}_{ij} - \mathbf{C}_{\text{gbl},ij} \begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mathbf{C}_{\text{grp},ij}^g \begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mathbf{C}_{\text{grp},ij}^h \begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} \right) \right\|^2 \right. \\ &\quad + \text{tr}(\mathbf{C}_{\text{gbl},ij}^T \mathbf{C}_{\text{gbl},ij} \Sigma_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})}) + \text{tr}((\mathbf{C}_{\text{grp},ij}^g)^T \mathbf{C}_{\text{grp},ij}^g \Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)}) + \text{tr}((\mathbf{C}_{\text{grp},ij}^h)^T \mathbf{C}_{\text{grp},ij}^h \Sigma_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)}) \\ &\quad + 2 \text{tr} \left[(\mathbf{C}_{\text{grp},ij}^g)^T \mathbf{C}_{\text{gbl},ij} E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right)^T \right\} \right] \\ &\quad + 2 \text{tr} \left[(\mathbf{C}_{\text{grp},ij}^h)^T \mathbf{C}_{\text{gbl},ij} E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \beta \\ \mathbf{u}_{\text{gbl}} \end{bmatrix} - \mu_{\mathbf{q}(\beta, \mathbf{u}_{\text{gbl}})} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\} \right] \\ &\quad \left. + 2 \text{tr} \left[(\mathbf{C}_{\text{grp},ij}^g)^T \mathbf{C}_{\text{grp},ij}^h E_{\mathbf{q}} \left\{ \left(\begin{bmatrix} \mathbf{u}_{\text{lin},i}^g \\ \mathbf{u}_{\text{grp},i}^g \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},i}^g, \mathbf{u}_{\text{grp},i}^g)} \right) \left(\begin{bmatrix} \mathbf{u}_{\text{lin},ij}^h \\ \mathbf{u}_{\text{grp},ij}^h \end{bmatrix} - \mu_{\mathbf{q}(\mathbf{u}_{\text{lin},ij}^h, \mathbf{u}_{\text{grp},ij}^h)} \right)^T \right\} \right] \right\} \end{aligned}$$

where $\mathbf{C}_{\text{gbl},ij} \equiv [\mathbf{X}_{ij} \mathbf{Z}_{\text{gbl},ij}]$, $\mathbf{C}_{\text{grp},ij}^g \equiv [\mathbf{X}_{ij} \mathbf{Z}_{\text{grp},ij}^g]$, $\mathbf{C}_{\text{grp},ij}^h \equiv [\mathbf{X}_{ij} \mathbf{Z}_{\text{grp},ij}^h]$ and with reciprocal moment $\mu_{\mathbf{q}(1/\sigma_\varepsilon^2)} = \xi_{\mathbf{q}(\sigma_\varepsilon^2)} / \lambda_{\mathbf{q}(\sigma_\varepsilon^2)}$,

$\mathbf{q}(\sigma_{\text{gbl}}^2)$ is an Inverse- χ^2 ($\xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)}, \lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}$) density function

where $\xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} = \nu_{\text{gbl}} + K_{\text{gbl}}$ and

$$\lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)} = \mu_{\mathbf{q}(1/a_{\text{gbl}})} + \|\mu_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}\|^2 + \text{tr}(\Sigma_{\mathbf{q}(\mathbf{u}_{\text{gbl}})}),$$

with reciprocal moment $\mu_{\mathbf{q}(1/\sigma_{\text{gbl}}^2)} = \xi_{\mathbf{q}(\sigma_{\text{gbl}}^2)} / \lambda_{\mathbf{q}(\sigma_{\text{gbl}}^2)}$,

$\mathbf{q}(\sigma_{\text{grp},g}^2)$ is an Inverse- χ^2 ($\xi_{\mathbf{q}(\sigma_{\text{grp},g}^2)}, \lambda_{\mathbf{q}(\sigma_{\text{grp},g}^2)}$) density function

where $\xi_{\mathbf{q}}(\sigma_{\text{grp},g}^2) = \nu_{\text{grp},g} + mK_{\text{grp}}^g$ and

$$\lambda_{\mathbf{q}}(\sigma_{\text{grp},g}^2) = \mu_{\mathbf{q}}(1/a_{\text{grp},g}) + \sum_{i=1}^m \left\{ \|\boldsymbol{\mu}_{\mathbf{q}}(\mathbf{u}_{\text{grp},i}^g)\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\mathbf{q}}(\mathbf{u}_{\text{grp},i}^g)) \right\},$$

with reciprocal moment $\mu_{\mathbf{q}}(1/\sigma_{\text{grp},g}^2) = \xi_{\mathbf{q}}(\sigma_{\text{grp},g}^2)/\lambda_{\mathbf{q}}(\sigma_{\text{grp},g}^2)$,

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

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

$$\boldsymbol{\Lambda}_{\mathbf{q}}(\boldsymbol{\Sigma}_g) = \mathbf{M}_{\mathbf{q}}(\mathbf{A}_{\boldsymbol{\Sigma}_g}^{-1}) + \sum_{i=1}^m \left(\boldsymbol{\mu}_{\mathbf{q}}(\mathbf{u}_{\text{lin},i}^g) \boldsymbol{\mu}_{\mathbf{q}}^T(\mathbf{u}_{\text{lin},i}^g) + \boldsymbol{\Sigma}_{\mathbf{q}}(\mathbf{u}_{\text{lin},i}^g) \right),$$

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

$\mathbf{q}(\sigma_{\text{grp},h}^2)$ is an Inverse- $\chi^2(\xi_{\mathbf{q}}(\sigma_{\text{grp},h}^2), \lambda_{\mathbf{q}}(\sigma_{\text{grp},h}^2))$ density function

where $\xi_{\mathbf{q}}(\sigma_{\text{grp},h}^2) = \nu_{\text{grp},h} + K_{\text{grp}}^h \sum_{i=1}^m n_i$ and

$$\lambda_{\mathbf{q}}(\sigma_{\text{grp},h}^2) = \mu_{\mathbf{q}}(1/a_{\text{grp},h}) + \sum_{i=1}^m \sum_{j=1}^{n_i} \left\{ \|\boldsymbol{\mu}_{\mathbf{q}}(\mathbf{u}_{\text{grp},ij}^h)\|^2 + \text{tr}(\boldsymbol{\Sigma}_{\mathbf{q}}(\mathbf{u}_{\text{grp},ij}^h)) \right\},$$

with reciprocal moment $\mu_{\mathbf{q}}(1/\sigma_{\text{grp},h}^2) = \xi_{\mathbf{q}}(\sigma_{\text{grp},h}^2)/\lambda_{\mathbf{q}}(\sigma_{\text{grp},h}^2)$,

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

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

$$\boldsymbol{\Lambda}_{\mathbf{q}}(\boldsymbol{\Sigma}_h) = \mathbf{M}_{\mathbf{q}}(\mathbf{A}_{\boldsymbol{\Sigma}_h}^{-1}) + \sum_{i=1}^m \sum_{j=1}^{n_i} \left(\boldsymbol{\mu}_{\mathbf{q}}(\mathbf{u}_{\text{lin},ij}^h) \boldsymbol{\mu}_{\mathbf{q}}^T(\mathbf{u}_{\text{lin},ij}^h) + \boldsymbol{\Sigma}_{\mathbf{q}}(\mathbf{u}_{\text{lin},ij}^h) \right),$$

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

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

where $\xi_{\mathbf{q}}(a_{\varepsilon}) = \nu_{\varepsilon} + 1$,

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

with reciprocal moment $\mu_{\mathbf{q}}(1/a_{\varepsilon}) = \xi_{\mathbf{q}}(a_{\varepsilon})/\lambda_{\mathbf{q}}(a_{\varepsilon})$,

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

where $\xi_{\mathbf{q}}(a_{\text{gbl}}) = \nu_{\text{gbl}} + 1$,

$$\lambda_{\mathbf{q}}(a_{\text{gbl}}) = \mu_{\mathbf{q}}(1/\sigma_{\text{gbl}}^2) + 1/(\nu_{\text{gbl}} s_{\text{gbl}}^2)$$

with reciprocal moment $\mu_{\mathbf{q}}(1/a_{\text{gbl}}) = \xi_{\mathbf{q}}(a_{\text{gbl}})/\lambda_{\mathbf{q}}(a_{\text{gbl}})$,

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

where $\xi_{\mathbf{q}}(a_{\text{grp},g}) = \nu_{\text{grp},g} + 1$,

$$\lambda_{\mathbf{q}}(a_{\text{grp},g}) = \mu_{\mathbf{q}}(1/\sigma_{\text{grp},g}^2) + 1/(\nu_{\text{grp},g} s_{\text{grp},g}^2)$$

with reciprocal moment $\mu_{q(1/a_{\text{grp},g})} = \xi_{q(a_{\text{grp},g})}/\lambda_{q(a_{\text{grp},g})}$ and

$q(\mathbf{A}_{\Sigma_g})$ is an Inverse-G-Wishart $\left(G_{\text{diag}}, \xi_{q(\mathbf{A}_{\Sigma_g})}, \Lambda_{q(\mathbf{A}_{\Sigma_g})}\right)$ density function

where $\xi_{q(\mathbf{A}_{\Sigma_g})} = \nu_{\Sigma_g} + 2$,

$$\Lambda_{q(\mathbf{A}_{\Sigma_g})} = \text{diag}\{\text{diagonal}(\mathbf{M}_{q(\Sigma_g^{-1})})\} + \Lambda_{\mathbf{A}_{\Sigma_g}}$$

with inverse moment $\mathbf{M}_{q(\mathbf{A}_{\Sigma_g}^{-1})} = \xi_{q(\mathbf{A}_{\Sigma_g})} \Lambda_{q(\mathbf{A}_{\Sigma_g})}^{-1}$,

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

where $\xi_{q(a_{\text{grp},h})} = \nu_{\text{grp},h} + 1$,

$$\lambda_{q(a_{\text{grp},h})} = \mu_{q(1/\sigma_{\text{grp},h}^2)} + 1/(\nu_{\text{grp},h} s_{\text{grp},h}^2)$$

with reciprocal moment $\mu_{q(1/a_{\text{grp},h})} = \xi_{q(a_{\text{grp},h})}/\lambda_{q(a_{\text{grp},h})}$ and

$q(\mathbf{A}_{\Sigma_h})$ is an Inverse-G-Wishart $\left(G_{\text{diag}}, \xi_{q(\mathbf{A}_{\Sigma_h})}, \Lambda_{q(\mathbf{A}_{\Sigma_h})}\right)$ density function

where $\xi_{q(\mathbf{A}_{\Sigma_h})} = \nu_{\Sigma_h} + 2$

$$\Lambda_{q(\mathbf{A}_{\Sigma_h})} = \text{diag}\{\text{diagonal}(\mathbf{M}_{q(\Sigma_h^{-1})})\} + \Lambda_{\mathbf{A}_{\Sigma_h}}$$

with inverse moment $\mathbf{M}_{q(\mathbf{A}_{\Sigma_h}^{-1})} = \xi_{q(\mathbf{A}_{\Sigma_h})} \Lambda_{q(\mathbf{A}_{\Sigma_h})}^{-1}$.

S.11 The SOLVETWOLEVELSPARSELEASTSQUARES Algorithm

The SOLVETWOLEVELSPARSELEASTSQUARES is listed in Nolan *et al.* (2018) and based on Theorem 2 of Nolan & Wand (2018). Given its centrality to Algorithms 1 and 2 we list it again here. The algorithm solves a sparse version of the the least squares problem:

$$\min_x \|b - Bx\|^2$$

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

$$\mathbf{B} \equiv \begin{bmatrix} B_1 & \dot{B}_1 & O & \cdots & O \\ B_2 & O & \dot{B}_2 & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ B_m & O & O & \cdots & \dot{B}_m \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \quad (\text{S.4})$$

The sub-matrices corresponding to the non-zero blocks of \mathbf{A} are labelled according to:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12,1} & \mathbf{A}^{12,2} & \cdots & \mathbf{A}^{12,m} \\ \mathbf{A}^{12,1T} & \mathbf{A}^{22,1} & \times & \cdots & \times \\ \mathbf{A}^{12,2T} & \times & \mathbf{A}^{22,2} & \cdots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}^{12,mT} & \times & \times & \cdots & \mathbf{A}^{22,m} \end{bmatrix}. \quad (\text{S.5})$$

with \times denoting sub-blocks that are not of interest. The SOLVETWOLEVELSPARSELEASTSQUARES algorithm is given in Algorithm S.1.

Algorithm S.1 SOLVETWOLEVELSPARSELEASTSQUARES for solving the two-level sparse matrix least squares problem: minimise $\|\mathbf{b} - \mathbf{B} \mathbf{x}\|^2$ in \mathbf{x} and sub-blocks of \mathbf{A}^{-1} corresponding to the non-zero sub-blocks of $\mathbf{A} = \mathbf{B}^T \mathbf{B}$. The sub-block notation is given by (S.4) and (S.5).

Inputs: $\{(\mathbf{b}_i(\tilde{n}_i \times 1), \mathbf{B}_i(\tilde{n}_i \times p), \dot{\mathbf{B}}_i(\tilde{n}_i \times q)) : 1 \leq i \leq m\}$

$\boldsymbol{\omega}_3 \leftarrow \text{NULL}$; $\boldsymbol{\Omega}_4 \leftarrow \text{NULL}$

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

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

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

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

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

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

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

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

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

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

Output: $(\mathbf{x}_1, \mathbf{A}^{11}, \{(\mathbf{x}_{2,i}, \mathbf{A}^{22,i}, \mathbf{A}^{12,i}) : 1 \leq i \leq m\})$

S.12 The SOLVETHREELEVELSPARSELEASTSQUARES Algorithm

The SOLVETHREELEVELSPARSELEASTSQUARES, listed in Nolan *et al.* (2018) is a proceduralization of Theorem 4 of Nolan & Wand (2018). Since it is central to Algorithms 3 and 4 we list it here. The SOLVETHREELEVELSPARSELEASTSQUARES algorithm is concerned with solving the sparse three-level version of

$$\min_x \|b - Bx\|^2$$

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

$$B \equiv \left[\text{stack}_{1 \leq i \leq m} \left\{ \text{stack}_{1 \leq j \leq n_i} (B_{ij}) \right\} \mid \text{blockdiag}_{1 \leq i \leq m} \left\{ \left[\text{stack}_{1 \leq j \leq n_i} (\dot{B}_{ij}) \mid \text{blockdiag}_{1 \leq j \leq n_i} (\ddot{B}_{ij}) \right] \right\} \right] \quad (\text{S.6})$$

and

$$b \equiv \text{stack}_{1 \leq i \leq m} \left\{ \text{stack}_{1 \leq j \leq n_i} (b_{ij}) \right\}. \quad (\text{S.7})$$

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

$$A^{-1} = \begin{bmatrix} A^{11} & A^{12,1} & A^{12,11} & A^{12,12} & A^{12,2} & A^{12,21} & A^{12,22} & A^{12,23} \\ A^{12,1T} & A^{22,1} & A^{12,1,1} & A^{12,1,2} & \times & \times & \times & \times \\ A^{12,11T} & A^{12,1,1T} & A^{22,11} & \times & \times & \times & \times & \times \\ A^{12,12T} & A^{12,1,2T} & \times & A^{22,12} & \times & \times & \times & \times \\ A^{12,2T} & \times & \times & \times & A^{22,2} & A^{12,2,1} & A^{12,2,2} & A^{12,2,3} \\ A^{12,21T} & \times & \times & \times & A^{12,2,1T} & A^{22,21} & \times & \times \\ A^{12,22T} & \times & \times & \times & A^{12,2,2T} & \times & A^{22,22} & \times \\ A^{12,23T} & \times & \times & \times & A^{12,2,3T} & \times & \times & A^{22,23} \end{bmatrix} \quad (\text{S.8})$$

for the $m = 2, n_1 = 2$ and $n_2 = 3$ case. The \times symbol denotes sub-blocks that are not of interest. The SOLVETHREELEVELSPARSELEASTSQUARES algorithm is given in Algorithm S.2.

Algorithm S.2 SOLVETHREELLEVELSPARSELEASTSQUARES for solving the three-level sparse matrix least squares problem: minimise $\|\mathbf{b} - \mathbf{B}\mathbf{x}\|^2$ in \mathbf{x} and sub-blocks of \mathbf{A}^{-1} corresponding to the non-zero sub-blocks of $\mathbf{A} = \mathbf{B}^T \mathbf{B}$. The sub-block notation is given by (S.6), (S.7) and (S.8).

Inputs: $\{(b_{ij}(\tilde{o}_{ij} \times 1), \mathbf{B}_{ij}(\tilde{o}_{ij} \times p), \dot{\mathbf{B}}_{ij}(\tilde{o}_{ij} \times q_1), \ddot{\mathbf{B}}_{ij}(\tilde{o}_{ij} \times q_2)) : 1 \leq i \leq m, 1 \leq j \leq n_i\}$

$\omega_7 \leftarrow \text{NULL}$; $\Omega_8 \leftarrow \text{NULL}$

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

$\omega_9 \leftarrow \text{NULL}$; $\Omega_{10} \leftarrow \text{NULL}$; $\Omega_{11} \leftarrow \text{NULL}$

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

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

$\mathbf{d}_{0ij} \leftarrow \mathbf{Q}_{ij}^T \mathbf{b}_{ij}$; $\mathbf{D}_{0ij} \leftarrow \mathbf{Q}_{ij}^T \mathbf{B}_{ij}$; $\dot{\mathbf{D}}_{0ij} \leftarrow \mathbf{Q}_{ij}^T \dot{\mathbf{B}}_{ij}$

$\mathbf{d}_{1ij} \leftarrow$ 1st q_2 rows of \mathbf{d}_{0ij} ; $\mathbf{d}_{2ij} \leftarrow$ remaining rows of \mathbf{d}_{0ij} ; $\omega_9 \leftarrow \begin{bmatrix} \omega_9 \\ \mathbf{d}_{2ij} \end{bmatrix}$

$\mathbf{D}_{1ij} \leftarrow$ 1st q_2 rows of \mathbf{D}_{0ij} ; $\mathbf{D}_{2ij} \leftarrow$ remaining rows of \mathbf{D}_{0ij} ; $\Omega_{10} \leftarrow \begin{bmatrix} \Omega_{10} \\ \mathbf{D}_{2ij} \end{bmatrix}$

$\dot{\mathbf{D}}_{1ij} \leftarrow$ 1st q_2 rows of $\dot{\mathbf{D}}_{0ij}$; $\dot{\mathbf{D}}_{2ij} \leftarrow$ remaining rows of $\dot{\mathbf{D}}_{0ij}$; $\Omega_{11} \leftarrow \begin{bmatrix} \Omega_{11} \\ \dot{\mathbf{D}}_{2ij} \end{bmatrix}$

Decompose $\Omega_{11} = \mathbf{Q}_i \begin{bmatrix} \mathbf{R}_i \\ \mathbf{0} \end{bmatrix}$ such that $\mathbf{Q}_i^{-1} = \mathbf{Q}_i^T$ and \mathbf{R}_i is upper-triangular.

$\mathbf{c}_{0i} \leftarrow \mathbf{Q}_i^T \omega_9$; $\mathbf{C}_{0i} \leftarrow \mathbf{Q}_i^T \Omega_{10}$

$\mathbf{c}_{1i} \leftarrow$ 1st q_1 rows of \mathbf{c}_{0i} ; $\mathbf{c}_{2i} \leftarrow$ remaining rows of \mathbf{c}_{0i} ; $\omega_7 \leftarrow \begin{bmatrix} \omega_7 \\ \mathbf{c}_{2i} \end{bmatrix}$

$\mathbf{C}_{1i} \leftarrow$ 1st q_1 rows of \mathbf{C}_{0i} ; $\mathbf{C}_{2i} \leftarrow$ remaining rows of \mathbf{C}_{0i} ; $\Omega_8 \leftarrow \begin{bmatrix} \Omega_8 \\ \mathbf{C}_{2i} \end{bmatrix}$

Decompose $\Omega_8 = \mathbf{Q} \begin{bmatrix} \mathbf{R} \\ \mathbf{0} \end{bmatrix}$ so that $\mathbf{Q}^{-1} = \mathbf{Q}^T$ and \mathbf{R} is upper-triangular.

$\mathbf{c} \leftarrow$ first p rows of $\mathbf{Q}^T \omega_7$; $\mathbf{x}_1 \leftarrow \mathbf{R}^{-1} \mathbf{c}$; $\mathbf{A}^{11} \leftarrow \mathbf{R}^{-1} \mathbf{R}^{-T}$

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

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

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

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

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

$\mathbf{A}^{12,ij} \leftarrow -\left\{ \mathbf{R}_{ij}^{-1} (\mathbf{D}_{1ij} \mathbf{A}^{11} + \dot{\mathbf{D}}_{1ij} \mathbf{A}^{12,iT}) \right\}^T$

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

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

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