Sparse Linear Mixed Model Selection via Streamlined Variational Bayes

Luca Maestrini

The Australian National University
luca.maestrini@anu.edu.au

joint work with

Emanuele Degani, *Dorota Toczydłowska* and *Matt Wand*

2022 ISBA World Meeting
Montreal, 26 June – 1 July 2022

Outlook

Introduction

- linear mixed models in general form
- two- and three-level models
- variational approximations

Efficient sparse linear mixed model selection

- streamlined variational inference
- global-local shrinkage priors

Simulated and real data experiments

- simulated data involving three-level models
- Perinatal data (two-level model)

Motivation

- A variety of applications (e.g. longitudinal or grouped data) can be studied through linear mixed-effects models having fixed and random effects in the linear predictor
- In some cases, datasets include a large number of predictors, but only a few are effectively relevant
- These predictors are often introduced as fixed effects in the model to guarantee parsimony and validity of the inferential conclusions, especially in sparse model settings
- A proper fixed effects selection procedure is recommended to identify relevant predictors
- Other approaches are based on
  - random effect selection induced by their covariance matrix decomposition (Chen and Dunson, 2003; Yang, 2013),
  - joint fixed and random effects selection (Kinney and Dunson, 2007; Yang *et al.*, 2020)

General Model Formulation

We study linear mixed models having the general form

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{u}|\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{G}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$
$$\sigma^2|a_{\sigma^2} \sim \text{Inverse-}\chi^2(\nu_{\sigma^2}, 1/a_{\sigma^2}), \quad a_{\sigma^2} \sim \text{Inverse-}\chi^2(1, 1/(\nu_{\sigma^2}s^2_{\sigma^2})),$$
$$\boldsymbol{G}|\boldsymbol{A}_{\boldsymbol{G}} \sim p(\boldsymbol{G}|\boldsymbol{A}_{\boldsymbol{G}}), \quad \boldsymbol{A}_{\boldsymbol{G}} \sim p(\boldsymbol{A}_{\boldsymbol{G}}),$$

where the auxiliary variable a_{σ^2} and matrix $\boldsymbol{A}_{\boldsymbol{G}}$ facilitate the implementation of variational inference (Maestrini and Wand, 2021)

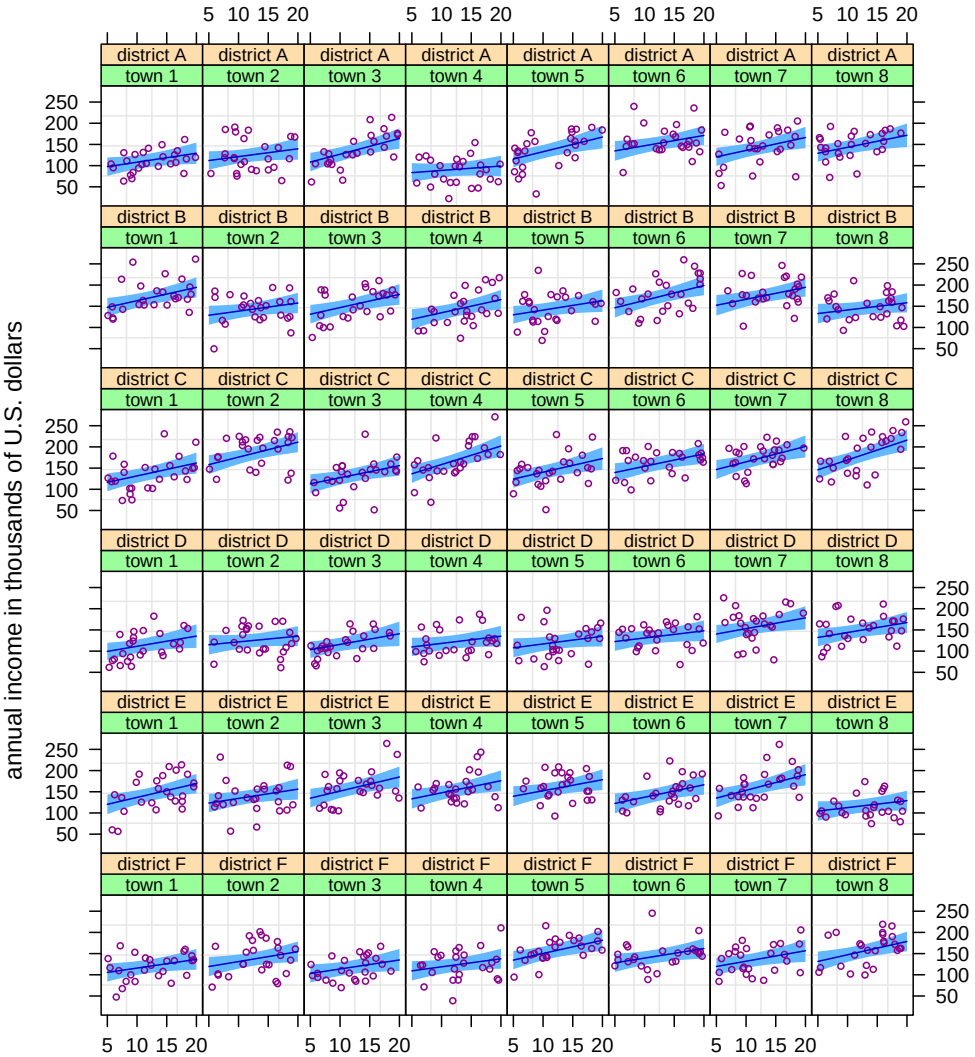
These are Gaussian response models where

- $\boldsymbol{\beta}$ is a vector of fixed effects
- $\boldsymbol{u}$ is a vector of random effects
- $\boldsymbol{X}$ is a fixed effects design matrix
- $\boldsymbol{Z}$ is a random effects design matrix
- $\boldsymbol{G}$ is a random effects covariance matrix

The focus of our work is on two- and three-level linear mixed models

Multilevel Data: Example

Three-level data with 6 districts, each having 8 towns where 25 residents are randomly selected (Nolan *et al.*, 2020)



Two- and Three-Level Models

A Two-level multilevel model for units clustered in m groups:

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

A three-level linear mixed model can be defined in terms of observations from the ith group and its jth subgroup as follows:

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

$$\begin{bmatrix} \boldsymbol{u}_i^{\mathrm{L1}} \\ \boldsymbol{u}_{ij}^{\mathrm{L2}} \end{bmatrix} \Big| \boldsymbol{\Sigma}^{\mathrm{L1}}, \boldsymbol{\Sigma}^{\mathrm{L2}} \overset{\text{ind}}{\sim} \mathrm{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}^{\mathrm{L1}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Sigma}^{\mathrm{L2}} \end{bmatrix} \right), \quad + \text{priors}$$

In this model

- $\boldsymbol{\Sigma}^{\mathrm{L1}}$ is the covariance matrix for the group-specific random effect vectors $\boldsymbol{u}_i^{\mathrm{L1}}$, $1 \leq i \leq m$
- $\boldsymbol{\Sigma}^{\mathrm{L2}}$ is the covariance matrix for the subgroup-specific random effect vectors $\boldsymbol{u}_{ij}^{\mathrm{L2}}$, $1 \leq i \leq m$, $1 \leq j \leq n_i$

Three-level Random Effects Models

The three-level model is a special case of the general specification

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{u}|\boldsymbol{G} \sim N(\boldsymbol{0}, \boldsymbol{G}), \quad + \text{priors}$$

with

$$\boldsymbol{y} = \underset{1\leq i\leq m}{\text{stack}} \left(\underset{1\leq j\leq n_i}{\text{stack}} (\boldsymbol{y}_{ij}) \right), \quad \boldsymbol{X} = \underset{1\leq i\leq m}{\text{stack}} \left(\underset{1\leq j\leq n_i}{\text{stack}} (\boldsymbol{X}_{ij}) \right),$$

$$\boldsymbol{Z} = \underset{1\leq i\leq m}{\text{blockdiag}} \left(\left[\begin{array}{c|c} \underset{1\leq j\leq n_i}{\text{stack}} (\boldsymbol{Z}_{ij}^{\text{L1}}) & \underset{1\leq j\leq n_i}{\text{blockdiag}} (\boldsymbol{Z}_{ij}^{\text{L2}}) \end{array} \right] \right),$$

$$\boldsymbol{u} = \underset{1\leq i\leq m}{\text{stack}} \left(\left[\begin{array}{c|c} (\boldsymbol{u}_{i}^{\text{L1}})^T & \left(\underset{1\leq j\leq n_i}{\text{stack}} (\boldsymbol{u}_{ij}^{\text{L2}}) \right)^T \end{array} \right]^T \right),$$

$$\boldsymbol{G} = \underset{1\leq i\leq m}{\text{blockdiag}} \left(\left[\begin{array}{cc} \boldsymbol{\Sigma}^{\text{L1}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{n_i} \otimes \boldsymbol{\Sigma}^{\text{L2}} \end{array} \right] \right)$$

Three-level Random Effects Models (3)

If, for example, the data is divided into $m = 3$ groups each having $n_1 = 2$, $n_2 = 3$, $n_3 = 2$ subgroups, then

$$
\boldsymbol{Z} = \begin{bmatrix}
\boldsymbol{Z}_{11}^{\mathrm{L1}} & \boldsymbol{Z}_{11}^{\mathrm{L2}} & & & & & & & & & & \\
\boldsymbol{Z}_{12}^{\mathrm{L1}} & & \boldsymbol{Z}_{12}^{\mathrm{L2}} & & & & & & & & & \\
& & & \boldsymbol{Z}_{21}^{\mathrm{L1}} & \boldsymbol{Z}_{21}^{\mathrm{L2}} & & & & & & & \\
& & & \boldsymbol{Z}_{22}^{\mathrm{L1}} & & \boldsymbol{Z}_{22}^{\mathrm{L2}} & & & & & & \\
& & & \boldsymbol{Z}_{23}^{\mathrm{L1}} & & & \boldsymbol{Z}_{23}^{\mathrm{L2}} & & & & & \\
& & & & & & & \boldsymbol{Z}_{31}^{\mathrm{L1}} & \boldsymbol{Z}_{31}^{\mathrm{L2}} & & \\
& & & & & & & \boldsymbol{Z}_{32}^{\mathrm{L1}} & & \boldsymbol{Z}_{32}^{\mathrm{L2}}
\end{bmatrix}
$$

Variational Approximations

Variational approximations are a class of techniques for making approximate model fitting and inference

Idea: Approximate a complex density function p with a simpler q

How to choose q?

Suppose p is a posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$. Then q is chosen minimize the Kullback–Leibler divergence (or other divergence measures):

- $KL(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\boldsymbol{y})) = \int q(\boldsymbol{\theta})\log\{q(\boldsymbol{\theta})/p(\boldsymbol{\theta}|\boldsymbol{y})\}\, d\boldsymbol{\theta}$
- $KL(p(\boldsymbol{\theta}|\boldsymbol{y})\|q(\boldsymbol{\theta})) = \int p(\boldsymbol{\theta}|\boldsymbol{y})\log\{p(\boldsymbol{\theta}|\boldsymbol{y})/q(\boldsymbol{\theta})\}\, d\boldsymbol{\theta}$
- etc...

Variational Bayes

The objective of variational Bayes approximations is to approximate the posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ with an approximating density q which is easier to obtain

- A mean field variational approximation $q^*(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the minimizer of the Kullback–Leibler divergence

$$\int q(\boldsymbol{\theta}) \log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{y})}\right\} d\boldsymbol{\theta} \quad \text{subject to} \quad q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q(\boldsymbol{\theta}_i),$$

  according to a partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ of $\boldsymbol{\theta}$

- If q belongs to an exponential family of distributions,

$$q^*(\boldsymbol{\theta}_i) \propto \exp\left\{E_{q(\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)} \log p(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i)\right\}, \qquad i = 1, \ldots, M,$$

  where $\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i$ are the entries of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_i$ omitted

MFVB Approximation

For the model in general form, a tractable variational approximation arises from

$$p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2, a_{\sigma^2}, \boldsymbol{G}, \boldsymbol{A_G}|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma^2) q(a_{\sigma^2})\, q(\boldsymbol{G})\, q(\boldsymbol{A_G})$$

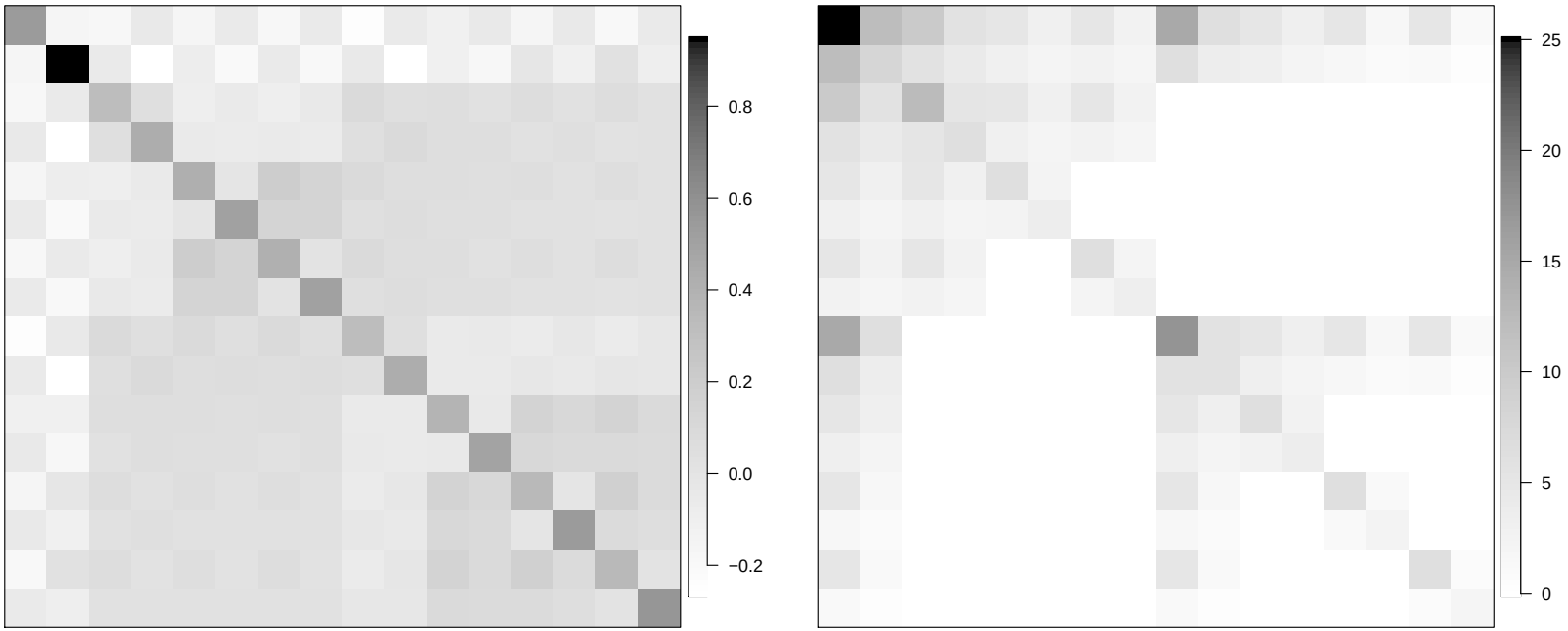
Under this restriction, the optimal MFVB approximating densities of the parameters of interest are the following:

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}) \text{ is a } \mathrm{N}\left(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right) \text{ density function,}$$
$$q^*(\sigma^2) \text{ is an Inverse-}\chi^2\left(\xi_{q(\sigma^2)}, \lambda_{q(\sigma^2)}\right) \text{ density function,}$$
$$q^*(a_{\sigma^2}) \text{ is an Inverse-}\chi^2\left(\xi_{q(a_{\sigma^2})}, \lambda_{q(a_{\sigma^2})}\right) \text{ density function,}$$
$$q^*(\boldsymbol{G}) \text{ and } q^*(\boldsymbol{A_G}) \text{ have form depending on the model specification}$$

Streamlined Variational Inference

- Note that $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ is $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$
- At each iteration of the variational algorithm, $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ has to be computed from a given $\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u})}$

$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ $\qquad$ $\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u})}$



Linear mixed model with three-level random effects, $m = 2$, $n_1 = 2$, $n_2 = 3$, $o_{ij} = 5$, $(i = 1, 2;\ 1 \leq j \leq n_1, n_2)$ and $p = q_1 = q_2 = 2$

Naïve MFVB

Let $\boldsymbol{C} = [\,\boldsymbol{X} \;\; \boldsymbol{Z}\,]$. The naïve MFVB algorithm updates for the parameters of $q(\boldsymbol{\beta}, \boldsymbol{u})$ and $q(\sigma^2)$ are:

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \left(\mu_{q(1/\sigma^2)}\, \boldsymbol{C}^T \boldsymbol{C} + \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & E_q(\boldsymbol{G}^{-1}) \end{bmatrix} \right)^{-1},$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \longleftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \left(\mu_{q(1/\sigma^2)}\, \boldsymbol{C}^T \boldsymbol{y} + \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \boldsymbol{0} \end{bmatrix} \right),$$

$$\xi_{q(\sigma^2)} \longleftarrow \nu_{\sigma^2} + n,$$

$$\lambda_{q(\sigma^2)} \longleftarrow \mu_{q(1/a_{\sigma^2})} + \|\boldsymbol{y} - \boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\|^2 + \operatorname{tr}\left\{ \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^T \boldsymbol{C} \right\}$$

The issues in large dimensions are:

- $\boldsymbol{Z}$ has a lot of zero elements that do not need to be stored
- $E_q(\boldsymbol{G}^{-1})$ becomes extremely sparse and inversion of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}^{-1}$ is unfeasible
- only few sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ are effectively useful for computing the other MFVB updates
- updating the expression for $\lambda_{q(\sigma^2)}$ requires operations involving huge sparse matrices

Streamlined MFVB

Streamlined MFVB:

- The mean field variational Bayes updates of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ and each of the sub-blocks of $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ that are relevant for variational inference are expressible as a sparse matrix problem of the form:

$$\boldsymbol{A}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} = \boldsymbol{a}$$

- The non-zero sub-blocks of $\boldsymbol{A}$ and $\boldsymbol{a}$ have close (but involved!) forms; see Nolan *et al.* (2020)
- Efficient matrix inversion and operations arise exploiting the sparse block-arrowhead matrix structure of $\boldsymbol{\Sigma}^{-1}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ that occurs when two- and three-level models are considered

Global-Local Shrinkage Priors

- We perform selection of fixed effects through priors on $\boldsymbol{\beta}$.
- Let $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^{\mathrm{R}}\boldsymbol{\beta}^{\mathrm{R}} + \boldsymbol{X}^{\mathrm{S}}\boldsymbol{\beta}^{\mathrm{S}} + \boldsymbol{X}^{\mathrm{A}}\boldsymbol{\beta}^{\mathrm{A}}$, where
  - $\boldsymbol{\beta}^{\mathrm{R}}$ are fixed effects sharing a predictor with a random effect
  - $\boldsymbol{\beta}^{\mathrm{S}}$ are fixed effects subject to selection
  - $\boldsymbol{\beta}^{\mathrm{A}}$ are other fixed effects
- We use priors

$$\boldsymbol{\beta}^{\mathrm{R}} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}^{\mathrm{R}}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{\mathrm{R}}}\right), \quad \boldsymbol{\beta}^{\mathrm{S}} \sim p(\boldsymbol{\beta}^{\mathrm{S}}) = \prod_{h=1}^{p_{\mathrm{S}}} p(\beta_h^{\mathrm{S}}), \quad \boldsymbol{\beta}^{\mathrm{A}} \sim \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}^{\mathrm{A}}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}^{\mathrm{A}}}\right)$$

  where $p(\beta_h^{\mathrm{S}})$ is the pdf of a global-local shrinkage prior
- These priors admit the scale mixture representation (Polson and Scott, 2011)

$$\beta_h^{\mathrm{S}} | \tau, \zeta_h \overset{\text{ind}}{\sim} \mathrm{N}(0, \tau^2/\zeta_h), \quad \tau \sim p(\tau), \quad \zeta_h \sim p(\zeta_h),$$

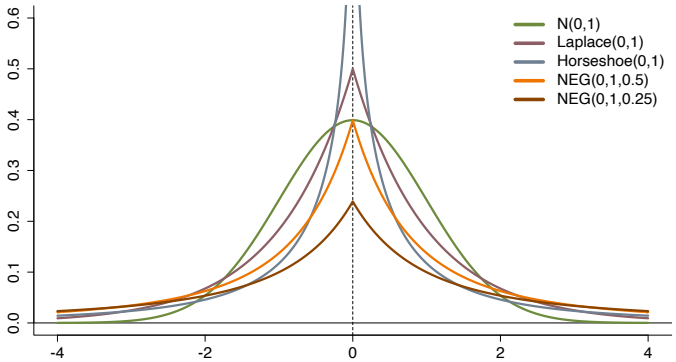
  where τ^2 is a global variance parameter, and ζ_h is a local which may induce a shrinkage effect towards zero to the hth fixed effect

Global-Local Shrinkage Priors and SAVS

- For the fixed effects subject to selection we consider the priors

$$\beta_h^S|\tau \overset{\text{iid}}{\sim} \text{Laplace}(0,\tau), \quad \beta_h^S|\tau \overset{\text{iid}}{\sim} \text{Horseshoe}(0,\tau), \quad \beta_h^S|\tau \overset{\text{iid}}{\sim} \text{NEG}(0,\tau;\lambda)$$

(NEG=Negative-Exponential-Gamma) and use the diffuse prior $\tau \sim \text{Half-Cauchy}(10^5)$



Global-local prior	$p(\beta_h^S\|\zeta_h,\tau)$	$p(\zeta_h\|a_{\zeta_h})$	$p(a_{\zeta_h})$
Laplace$(0,\tau)$	$N(0,\tau^2/\zeta_h)$	Inverse-$\chi^2(2,1)$	–
Horseshoe$(0,\tau)$	$N(0,\tau^2/\zeta_h)$	Gamma$(1/2, a_{\zeta_j})$	Gamma$(1/2,1)$
NEG$(0,\tau;\lambda)$	$N(0,\tau^2/\zeta_h)$	Inverse-$\chi^2(2, 2a_{\zeta_j})$	Gamma$(\lambda,1)$

- The posterior densities of irrelevant β_h^s's are shrunk towards zero but not sparsified
- We perform fixed-effects selection via the signal adaptive variable selector (SAVS) of Ray and Bhattacharya (2018), which is hyperparameters-tuning free:

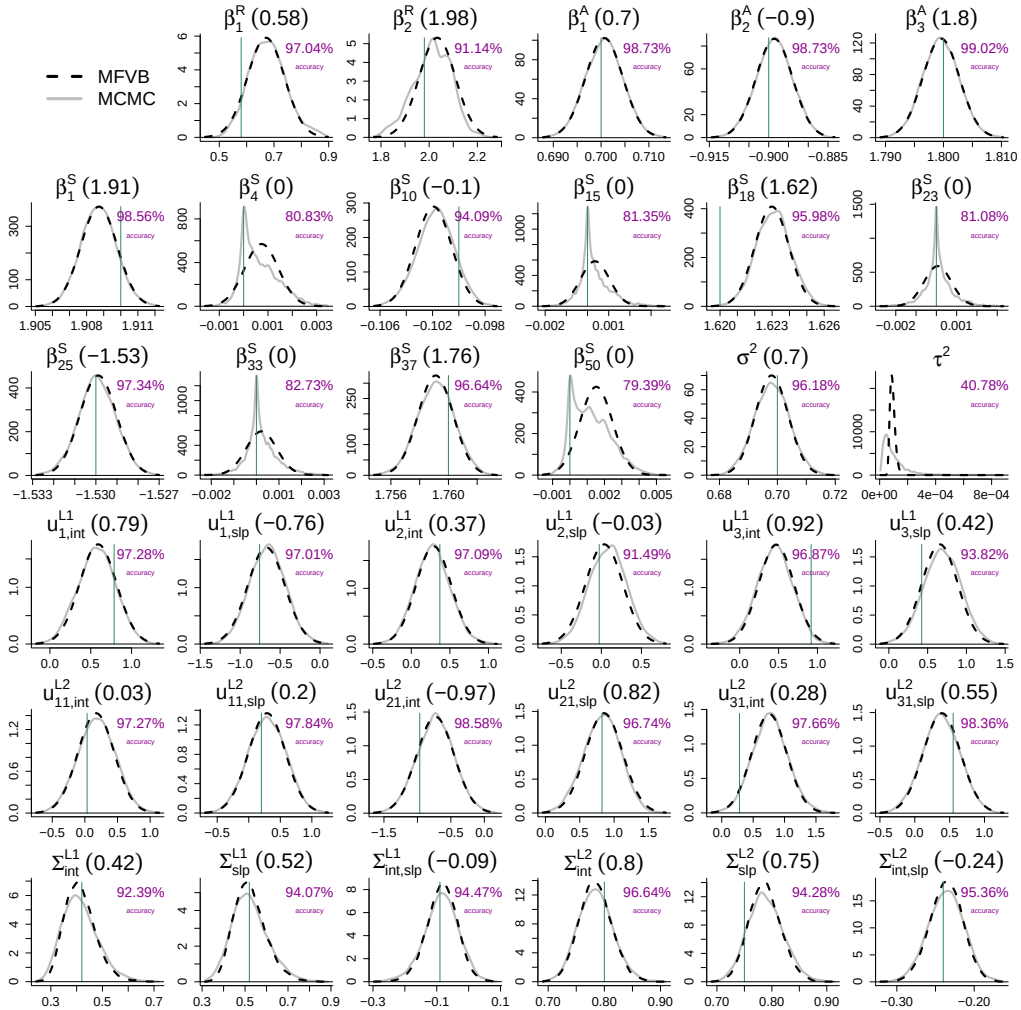
$$\text{if } \|\boldsymbol{x}_h\|^2 \leq \left|\mu^*_{q(\beta_h^S)}\right|^{-3}, \text{ then } \beta_h^S \text{ is discarded}$$

Simulated Data: Settings

- 50 simulated datasets from a three-level linear mixed model with $m = 100$ groups, each with $n_i = 15$ sub-groups of $o_{ij} = 20$ units each
- Random intercept and slope for both the group and sub-group levels ($q^{\mathrm{L1}} = q^{\mathrm{L2}} = p_{\mathrm{R}} = 2$), and $p_{\mathrm{A}} = 3$ additional fixed effects not subject to selection
- $p_{\mathrm{S}} = 50$ fixed effects subject to selection, where 40 of them are irrelevant (true value equal to zero); $\beta^{\mathrm{S}}_h|\tau \overset{\text{iid}}{\sim} \text{Horseshoe}(0, \tau)$, for $1 \leq h \leq 50$
- Hence, $\boldsymbol{X}$ has size 30,000 $\times$ 55, and $\boldsymbol{Z}$ has size 30,000 $\times$ 3,200 (96*10^6 cells, of which 99.875% are empty and do not contain any data)
- Diffuse priors for all the model parameters
- Both the naïve and streamlined MFVB approximations are obtained (200 iterations)
- Comparison with MCMC: Gibbs sampler (5,000 burn-in, 25,000 kept, thinning of size 5)
- All algorithms are implemented in C++

Simulated Data: Approximation Assessment

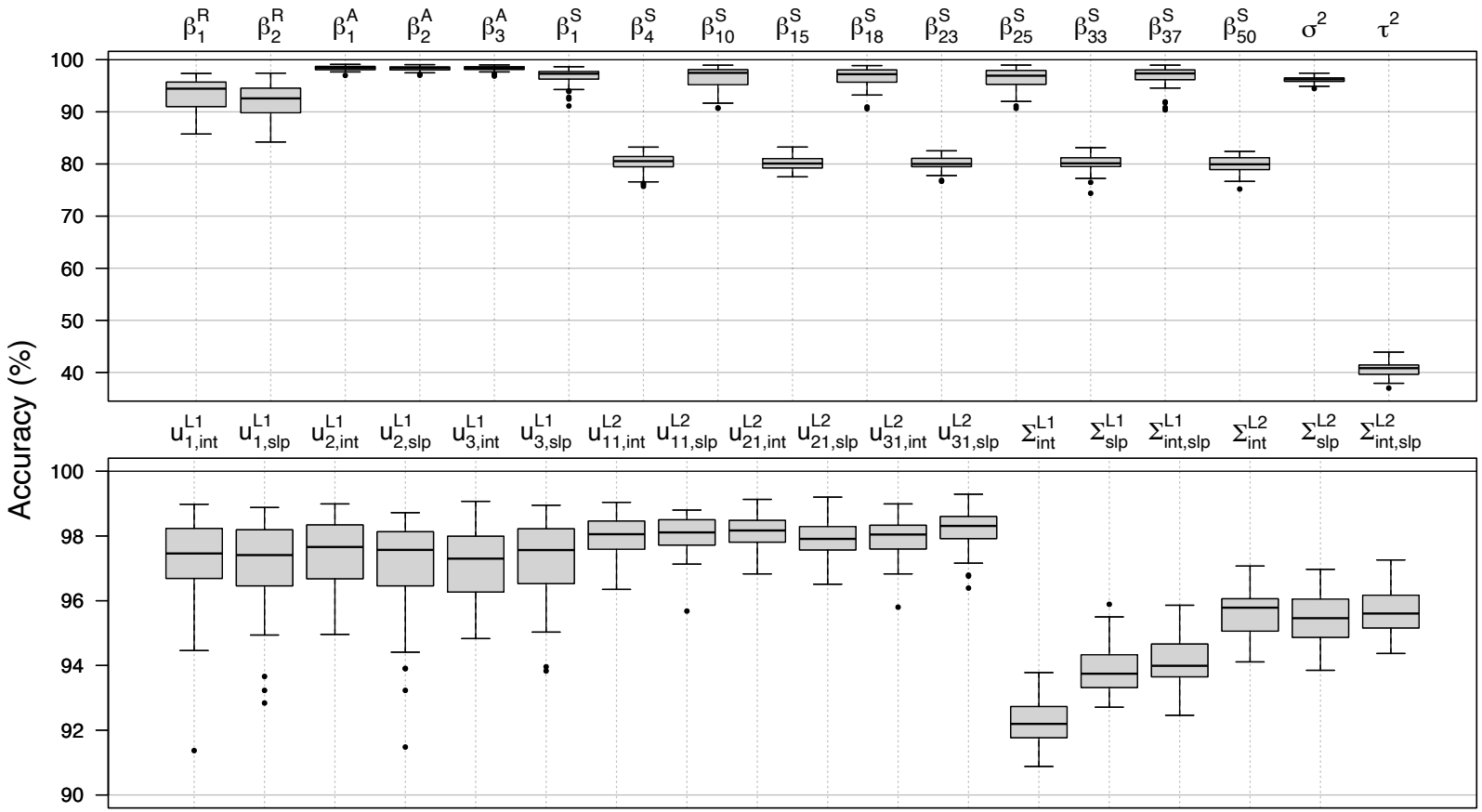
Optimal $q^*(\theta)$ densities obtained via MFVB (black dashed curves) and the corresponding MCMC-based $p(\theta|\boldsymbol{y})$ densities (grey curves). Vertical lines indicate the parameters true values.



Simulated Data: Accuracy Assessment

Boxplots of the accuracy scores for a selection of model parameters and random effects

$\text{Accuracy}\{q^*(\theta)\} \equiv (1 - \frac{1}{2}\int |q^*(\boldsymbol{\theta}) - p(\theta|\boldsymbol{y})|\, \mathrm{d}\theta) \times 100\%$



Simulated Data: Selection Performance

- Get MFVB approximations for each of the 50 datasets and priors

$$\beta_h^S \overset{\text{iid}}{\sim} \text{N}(0, 10^{10}), \quad \beta_h^S|\tau \overset{\text{iid}}{\sim} \text{Laplace}(0, \tau),$$

$$\beta_h^S|\tau \overset{\text{iid}}{\sim} \text{Horseshoe}(0, \tau), \quad \beta_h^S|\tau \overset{\text{iid}}{\sim} \text{NEG}(0, \tau; 0.25)$$

- After sparsifying the $q^*(\beta_h^S)$ via the SAVS procedure, the selection performance is measured using the F_1-score:

$$F_1 = \left(\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \right) \times 100\%, \ \text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \ \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

TP $=$ # of fixed effects with true $\beta_h^S \neq 0$ and selected by SAVS,

TN $=$ # of fixed effects with true $\beta_h^S = 0$ and *not* selected by SAVS,

FP $=$ # of fixed effects with true $\beta_h^S = 0$ and selected by SAVS,

FN $=$ # of fixed effects with true $\beta_h^S \neq 0$ and *not* selected by SAVS

- Selection performance:
  - Gaussian: median F_1-score 63.5% (q1: 43.5% – q3: 94.2%)
  - Laplace: 95.24% (80%–100%)
  - Horseshoe and NEG: 100% (100%–100%, i.e. perfect fixed effects selection for each of the 50 datasets)

Simulated Data: Speed and Memory Saving

Speed and memory saving assessment for:

- three-level model
- p_S fixed effects subject to selection in $\{25, 100, 200\}$
- number of groups m in $\{10, 50, 100, 200\}$
- n_i and o_{ij} uniformly from the discrete sets $\{10, \ldots, 20\}$ and $\{20, \ldots, 30\}$

The table displays average and standard deviation (in brackets) of

- running time (seconds)
- total size of required data input (megabytes)

		Total runtime of the algorithm (seconds)				Total size of the required data inputs (megabytes)		
m	p_S	Streamlined MFVB	Naïve MFVB	$\frac{\text{Naïve MFVB}}{\text{Streamlined MFVB}}$	MCMC	Streamlined MFVB	Naïve MFVB	$\frac{\text{Naïve MFVB}}{\text{Streamlined MFVB}}$
10	25	$0.70_{(0.06)}$	$3.86_{(0.50)}$	**5.54**	$13.03_{(0.96)}$	$1.39_{(0.07)}$	$11.06_{(1.00)}$	**7.97**
	100	$3.90_{(0.23)}$	$6.51_{(0.82)}$	**1.67**	$46.99_{(1.47)}$	$3.68_{(0.24)}$	$12.73_{(1.34)}$	**3.46**
	200	$9.71_{(1.22)}$	$12.09_{(2.04)}$	**1.24**	$176.22_{(9.37)}$	$6.60_{(0.52)}$	$15.09_{(1.89)}$	**2.29**
50	25	$3.29_{(0.11)}$	$365.07_{(37.10)}$	**111.01**	$63.28_{(3.76)}$	$6.69_{(0.20)}$	$240.40_{(13.92)}$	**35.91**
	100	$19.66_{(0.87)}$	$415.29_{(51.66)}$	**21.12**	$138.42_{(5.72)}$	$18.37_{(0.81)}$	$254.04_{(20.84)}$	**13.83**
	200	$49.25_{(1.47)}$	$501.33_{(39.76)}$	**10.18**	$305.25_{(6.51)}$	$34.00_{(1.04)}$	$269.34_{(14.90)}$	**7.92**
100	25	$6.77_{(0.32)}$	$2877.33_{(177.77)}$	**424.72**	$151.05_{(10.48)}$	$13.56_{(0.34)}$	$967.34_{(42.15)}$	**71.33**
	100	$40.42_{(1.66)}$	$3172.66_{(116.83)}$	**78.50**	$266.71_{(10.06)}$	$36.77_{(1.29)}$	$1010.25_{(26.94)}$	**27.47**
	200	$97.81_{(2.07)}$	$3403.55_{(204.38)}$	**34.80**	$494.15_{(9.34)}$	$67.53_{(1.31)}$	$1013.90_{(48.67)}$	**15.01**
200	25	$13.30_{(0.30)}$	$>$ 5 hours	$>$ **1355**	$328.87_{(20.04)}$	$26.90_{(0.56)}$	$3817.06_{(113.85)}$	**141.88**
	100	$79.97_{(1.18)}$	$>$ 5 hours	$>$ **225**	$578.88_{(13.45)}$	$73.74_{(1.01)}$	$3941.15_{(97.90)}$	**53.45**
	200	$197.20_{(4.13)}$	$>$ 5 hours	$>$ **95**	$904.64_{(15.77)}$	$135.84_{(2.73)}$	$3991.66_{(102.48)}$	**29.38**

Perinatal Data (1)

National Collaborative Perinatal Project data (Klebanoff, 2009)

- A multisite prospective cohort study which took place in the United States of America between 1959 and 1974.
- Designed to identify the effects of complications during pregnancy or the perinatal period on birth and child outcomes
- Focus on predicting the height-for-age z-score (standardized measure of the WHO for the height of children after accounting for age) for 37,257 infants followed longitudinally over their first year of life

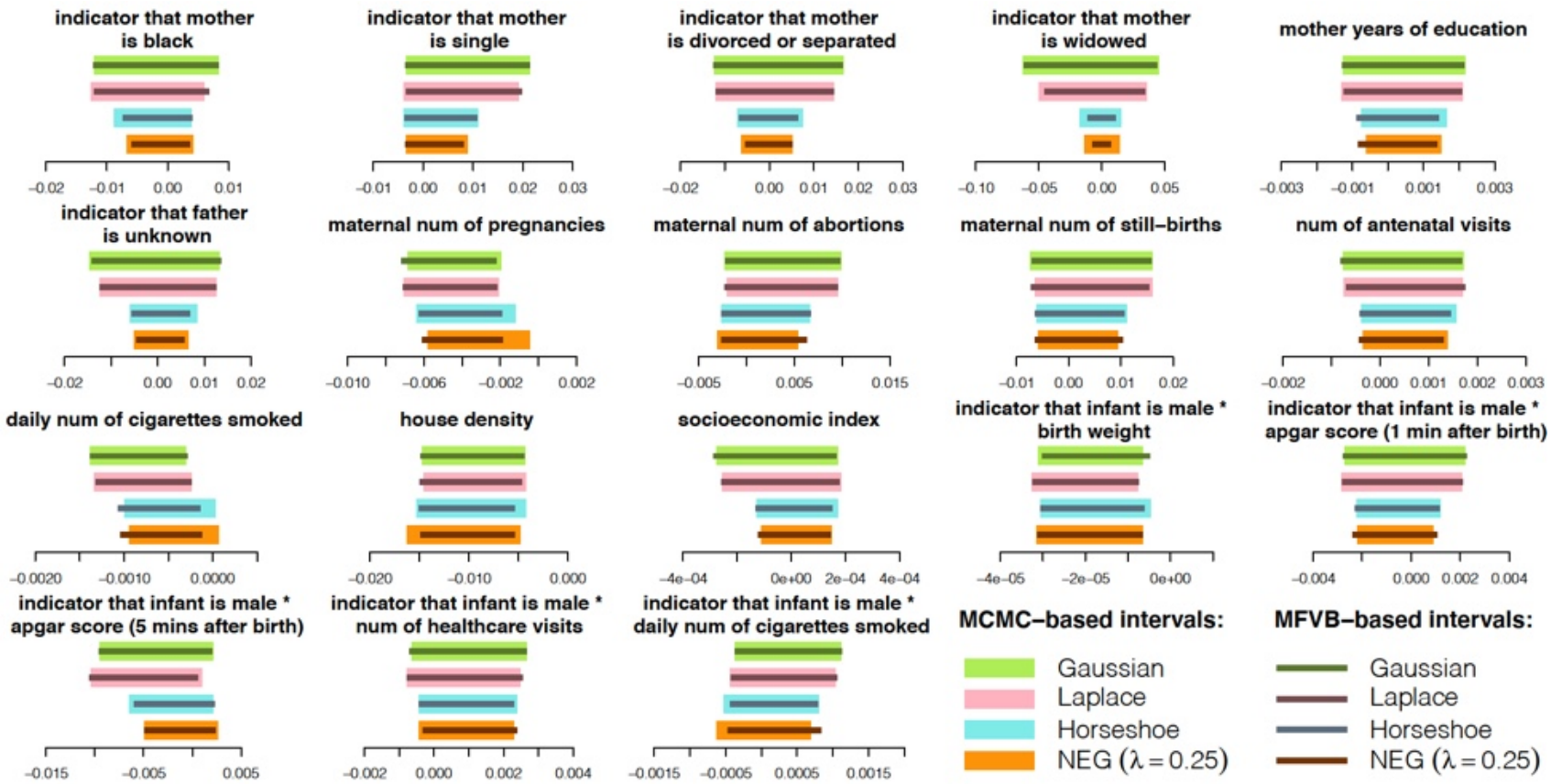
Perinatal Data (1)

National Collaborative Perinatal Project data (Klebanoff, 2009)

- A multisite prospective cohort study which took place in the United States of America between 1959 and 1974.
- Designed to identify the effects of complications during pregnancy or the perinatal period on birth and child outcomes
- Focus on predicting the height-for-age z-score (standardized measure of the WHO for the height of children after accounting for age) for 37,257 infants followed longitudinally over their first year of life
- We fit a two-level linear mixed model with
  - Intercept, age of the infant and its square, plus 38 candidate predictors subject to selection as fixed effects
  - Intercept, age of the infant and its square as random effects
  - Gaussian, Laplace, Horseshoe and Negative-Exponential-Gamma priors for fixed effects subject to selection
- MCMC (5,000 burn-in iterations followed by 25,000 iterations with thinning factor of 5) took more than 35 minutes to run
- Streamlined MFVB (200 iterations) took around 2 minutes to run

Perinatal Data (2)

The 90% high posterior density credible intervals for some of the fixed effects subject to selection



Extensions

More accurate variational approximations can be obtained at higher computational cost when using

- global-local prior specifications that are not based on auxiliary variables (see Neville *et al.*, 2014)
- less restrictive mean-field factorizations on the fixed effects, e.g. $q(\boldsymbol{\beta}, \boldsymbol{u}) = q(\boldsymbol{\beta}^{\mathrm{R}}, \boldsymbol{u})q(\boldsymbol{\beta}^{\mathrm{A}}, \boldsymbol{\beta}^{\mathrm{S}})$

With relatively little changes in the streamlined MFVB algorithms, it is possible to treat models with

- other response types (binary data with logit/probit link, Poisson, t, Skew Normal, Skew t, etc...)
- other types of priors for fixed effects selection (e.g. Bayesian lasso and spike-and-slab priors)

Other extensions include the development of streamlined variational inference for models with

- unit-specific errors, heteroskedastic covariance structures for groups and sub-groups, higher levels of nesting or crossed random effects
- priors for selecting random effects (however, this would entail formulating new streamlining strategies for the variational algorithms)

References

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690–698.

Degani, E., Maestrini, L., Toczydłowska, D. and Wand, M. P. (2022+). Sparse linear mixed model selection via streamlined variational Bayes. *arXiv:2110.07048*

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

Maestrini, L. and Wand, M.P. (2021). The Inverse G-Wishart Distribution and Variational Message Passing. *Australian and New Zealand Journal of Statistics*, **63**, 517–541

Neville, S. E., Ormerod, J. T and Wand, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, **8**, 1113–1151

Nolan, T.H., Menictas, M. and Wand, M.P. (2020). Streamlined computing for variational inference with higher level random effects. *Journal of Machine Learning Research*, **21**, 1–62

Polson, N. G. and Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, **9**, 501–538

Yang, M. (2013). Bayesian nonparametric centered random effects models with variable selection. *Biometrical Journal*, **55**, 217–230

Ray, P. and Bhattacharya, A. (2018). Signal adaptive variable selector for the horseshoe prior. *arXiv:1810.09004*

Yang, M., Wang, M., and Dong, G. (2020). Bayesian variable selection for mixed effects model with shrinkage prior. *Computational Statistics*, **35**, 227–243

Contacts

E-mail: *luca.maestrini@anu.edu.au*

Personal website: https://sites.google.com/view/lucamaestrini

Twitter: LucaMaeStats

Article:
Degani, E., Maestrini, L., Toczydłowska, D. and Wand, M. P. (2022+). Sparse linear mixed model selection via streamlined variational Bayes. *arXiv:2110.07048*