

Sparse Variational Analysis of Large Longitudinal Data Sets

Artin Armagan*

Department of Statistical Science, Duke University, Durham, NC 27708

David Dunson

Department of Statistical Science, Duke University, Durham, NC 27708

Abstract

It is increasingly common to be faced with longitudinal or multi-level data sets that have large number of predictors and/or a large sample size. Current methods of fitting and inference for mixed effects models tend to perform poorly in such settings. When there are many variables, it is appealing to allow uncertainty in subset selection and to obtain a sparse characterization of the data. Bayesian methods are available to address these goals using Markov chain Monte Carlo (MCMC), but MCMC is very computationally expensive and can be infeasible in large p and/or large n problems. As a fast approximate Bayes solution, we recommend a novel approximation to the posterior relying on variational methods. Variational methods are used to approximate the posterior of the parameters in a decomposition of the variance components, with priors chosen to obtain a sparse solution that allows selection of random effects. The method is evaluated through a simulation study, and applied to an epidemiological application.

Key words: Mixed-effects model, Variational approximations, Shrinkage estimation

*Corresponding author

Email addresses: `artin@stat.duke.edu` (Artin Armagan), `dunson@stat.duke.edu` (David Dunson)

1. Introduction

It is often of interest to fit a hierarchical model in settings involving large numbers of predictors (p) and/or large sample size (n). For example, in a large prospective epidemiology study, one may obtain longitudinal data for tens of thousands of subjects, while also collecting ~ 100 predictors. Even in more modest studies, involving thousands of subjects, the number of predictors collected is often large. Unfortunately, current methods for inference in mixed effects models are not designed to accommodate large p and/or large n . This article proposes a method for obtaining sparse approximate Bayes inferences in such problems using variational methods (Jordan et al., 1999; Jaakkola and Jordan, 2000).

For concreteness we focus on the linear mixed effects (LME) model (Laird and Ware, 1982), though the proposed methods can be applied directly in many other hierarchical models. When considering LMEs in settings involving moderate to large p , it is appealing to consider methods that encourage sparse estimation of the random effects covariance matrix. There are a variety of methods available in the literature, including approaches based on Bayesian methods implemented with MCMC (Chen and Dunson, 2003; Kinney and Dunson, 2007; Frühwirth-Schnatter and Tüchler, 2008) and methods based on fast shrinkage estimation (Foster et al., 2007).

Frequentist procedures encounter convergence problems (Pennell and Dunson, 2007) and MCMC based methods tend to be computationally intensive and not to scale well as p and/or n increases. The methods relying on stochastic search variable selection (SSVS) algorithms (George and McCulloch, 1997) face difficulties when p increases beyond ~ 30 in linear regression applications, with the computational burden substantially greater in hierarchical models involving random effects selection. Approaches have been proposed to make MCMC implementations of hierarchical models feasible in large data sets (Huang and Gelman, 2008; Pennell and Dunson, 2007). However, these approaches do not solve the large p problem or allow sparse estimation or selection of random effects covariances. In addition, the algorithms are still time consuming to implement.

Model selection through shrinkage estimation has gained much popularity since the Lasso of (Tibshirani, 1996). Similar shrinkage effects were later obtained through hierarchical modeling of the regression coefficients in the Bayesian paradigm. A few examples of these are (Tipping, 2001; Bishop and Tipping, 2000; Figueiredo, 2003; Park and Casella, 2008). Most approaches have relied on *maximum a posteriori* (MAP) estimation. MAP estimation produces a sparse point estimate with no measure of uncertainty, motivating MCMC and variational

methods.

It would be appealing to have a fast approach that could be implemented much more rapidly in cases involving moderate to large data sets and numbers of variables, while producing sparse estimates and allowing approximate Bayesian inferences on predictor effects. In particular, it would be very appealing to have an approximation to the marginal posteriors instead of simply obtaining a point estimate. Basing inferences on point estimates does not account for uncertainty in the estimation process, and hence is not useful in applications, such as epidemiology.

One possibility is to rely on a variational approximation to the posterior distribution (Jordan et al., 1999; Jaakkola and Jordan, 2000; Bishop and Tipping, 2000). Within this framework, we develop a method for sparse covariance estimation relying on a decomposition and the use of heavy-tailed priors in a related manner to (Tipping, 2001), though they did not consider estimation of covariance matrices.

Section 2 reviews the variational methods. Section 3 proposes a shrinkage model to encourage sparse estimates, inducing variable selection and gives the variational approximations to the posteriors. Section 4 presents a simulation study to assess the performance of the proposed methods, Section 5 applies the method to a large epidemiologic study of child growth, and finally Section 6 discusses the results.

2. Variational inference

Except in very simple conjugate models, the marginal likelihood of the data is not available analytically. As an alternative to MCMC and Laplace approximations (Tierney and Kadane, 1986), a lower-bound on marginal likelihoods may be obtained via variational methods (Jordan et al., 1999) yielding approximate posterior distributions on the model parameters. Let θ be the vector of all unobserved quantities in the model and \mathbf{y} be the observed data. Given a distribution $q(\theta)$, the marginal log-likelihood can be decomposed as

$$\log p(\mathbf{y}) = \underbrace{\int q(\theta) \log \frac{p(\mathbf{y}, \theta)}{q(\theta)} d\theta}_{\mathcal{L}} + KL(q||p), \tag{1}$$

where $p(\mathbf{y}, \theta)$ is an unnormalized posterior density of θ and $KL(.||.)$ denotes the Kullback-Leibler divergence between two distributions. Since this quantity is a strictly non-negative one and is equal to 0 only when $p(\theta|\mathbf{y}) = q(\theta)$, the first term

in (1) constitutes a lower-bound on $\log p(\mathbf{y})$. It is evident that maximizing the first term in the right hand side of (1) is equivalent to minimizing the second term in the right hand side, suggesting that $q(\boldsymbol{\theta})$ is an approximation to the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$.

Following (Bishop and Tipping, 2000) we consider a factorized form

$$q(\boldsymbol{\theta}) = \prod_i q_i(\theta_i), \quad (2)$$

where θ_i is a sub-vector of $\boldsymbol{\theta}$ and there are no restrictions on the form of $q_i(\theta_i)$. Then the lower-bound can be maximized with respect to $q_i(\theta_i)$ yielding the solution

$$q_i(\theta_i) = \frac{\exp\langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{\theta_{j \neq i}}}{\int \exp\langle \log p(\mathbf{y}, \boldsymbol{\theta}) \rangle_{\theta_{j \neq i}} d\theta_i}, \quad (3)$$

where $\langle \cdot \rangle_{\theta_{j \neq i}}$ denotes the expectation with respect to the distributions $q_j(\theta_j)$ for $j \neq i$. As we will see, due to conjugacy we will obtain for our model, these expectations will be easily evaluated yielding standard distributions for $q_i(\theta_i)$ with parameters expressed in terms of the moments of θ_i . Thus the procedure will consist of initializing the expectations required and re-iterating through them updating the expectations with respect to the densities provided by (3).

3. The Model

3.1. The Standard Model

Suppose there are n subjects under study, with n_i observations for the i th subject. For subject i at observation j , let y_{ij} denote the response, let \mathbf{x}_{ij} and \mathbf{z}_{ij} denote $p \times 1$ and $q \times 1$ vectors for predictors. Then, a linear mixed effects model can be written as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\alpha} + \mathbf{Z}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i, \quad (4)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$, $\mathbf{X}_i = (\mathbf{x}'_{i1}, \dots, \mathbf{x}'_{in_i})'$, $\mathbf{Z}_i = (\mathbf{z}'_{i1}, \dots, \mathbf{z}'_{in_i})'$, $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unknown fixed effects, $\boldsymbol{\beta}_i$ is a $q \times 1$ vector unknown subject-specific random effects with $\boldsymbol{\beta}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, and the elements of the residual vector, $\boldsymbol{\varepsilon}_i$, are $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

Given the formulation in (4), the joint density of the observations given the model parameters can be written as

$$p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\sum_{i=1}^n n_i/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\alpha} - \mathbf{z}'_{ij}\boldsymbol{\beta}_i)^2\right\}. \quad (5)$$

From the model definition we know that the random effects are distributed as $\mathcal{N}(\mathbf{0}, \mathbf{D})$. Here, once the appropriate priors are placed on the model parameters, the inference is straight-forward both in exact (via MCMC) and approximate (via variational methods) cases. To save space, the variational approximations to the posteriors of the model parameters will be given only for the shrinkage model explained in the following section and can easily be reduced to the standard model case.

3.2. The Shrinkage Model

Let $\mathbf{D} = \mathbf{\Lambda}\mathbf{B}\mathbf{\Lambda}$ where \mathbf{B} is a symmetric, positive-definite matrix and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_q)$ with $\lambda_k \in \mathbb{R}$ with no further restrictions. This decomposition is not unique yet is sufficient to guarantee the positive-semidefiniteness of \mathbf{D} . This can easily be verified. Let $\mathbf{B} = \mathbf{\Gamma}\mathbf{\Gamma}'$ be a Cholesky decomposition for \mathbf{B} . Also let $\mathbf{\Lambda}\mathbf{\Gamma} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ be a singular value decomposition for $\mathbf{\Lambda}\mathbf{\Gamma}$. Then $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'\mathbf{V}\mathbf{\Sigma}\mathbf{U}' = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}\mathbf{U}'$ is an eigenvalue decomposition for \mathbf{D} where the eigenvalues are nonnegative.

Let us re-write (4) as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{Z}_i\mathbf{\Lambda}\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (6)$$

where $\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{B})$ and λ_k acts as a scaling factor on the k th row and column of the random effects covariance \mathbf{D} . Although the parameterization is redundant, it has been noted that the redundant parameterizations are often useful for computational reasons and for inducing new classes of priors with appealing properties (Gelman, 2006). The incorporation of λ_k allows greater control on adaptive predictor-dependent shrinkage, with values $\lambda_k \approx 0$ (along with small corresponding diagonals in \mathbf{B}) leading to the k th predictor being effectively excluded from the random effects component of the model through setting the values in the k th row and column of \mathbf{D} close to zero. This maintains the positive-semidefinite constraint. One issue with redundant parameterization is the lack of identifiability in a frequentist sense, i.e. it will lead to a likelihood which comprises multiple ridges along possible combinations of $\mathbf{\Lambda}$ and \mathbf{B} . This does not create insurmountable difficulties for Bayesian procedures as a non-flat prior should take care of this problem. When MCMC is used, the sampling of λ_k and \mathbf{B} would occur along these ridges where the prior assigns a positive density. Multiple modes will exist due to the fact that each λ_k may take either sign.

The variational procedure used will converge to one of multiple exchangeable modes which live on the aforementioned ridges in the posterior. The tracking of the lower-bound plays an important role to stop the iterative procedure. As the

lower-bound stops its monotonic increase (according to some preset criterion), we stop the procedure and assume that any further change in λ_k and \mathbf{B} will not change inferences as we are moving along one of these ridges.

Given the formulation in (6), the joint density of the observations given the model parameters can be written as

$$p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\Lambda}, \mathbf{b}, \sigma^2) = (2\pi\sigma^2)^{-\sum_{i=1}^n \frac{n_i}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\alpha} - \mathbf{z}'_{ij}\boldsymbol{\Lambda}\mathbf{b}_i)^2\right\}. \quad (7)$$

From the model definition we know that the \mathbf{b}_i is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{B})$.

3.2.1. Priors and Posteriors

After this decomposition the joint (conditional) density of the observations remains almost identical, replacing $\boldsymbol{\beta}$ by $\boldsymbol{\Lambda}\mathbf{b}_i$ in (7). Notice that λ_k and \mathbf{b}_{ik} are interchangeable which is going to allow us to model λ_k as *redundant* random-effects coefficients.

We will use independent t priors for α_k and λ_k due to its shrinkage reinforcing quality. This will be accomplished through the scale mixtures of normals (West, 1987) due to the conjugacy properties, i.e. $\alpha_k \sim \mathcal{N}(0, a_k)$ and $\lambda_k \sim \mathcal{N}(0, v_k)$ where $a_k^{-1}, v_k^{-1} \sim \mathcal{G}(\eta_0, \zeta_0)$. Under this setup, we would hope that those α_k and λ_k corresponding to insignificant fixed and random effects would shrink towards the neighborhood of zero. This will allow us to obtain a much smaller set of fixed effects for prediction purposes as well as a much more compact covariance structure on the random-effects coefficients. The usefulness of this approach will be especially emphasized in high dimensional problems. We also set $\sigma^2 \sim \mathcal{IG}(c_0, d_0)$ and $\mathbf{B} \sim \mathcal{IW}(n_0, \boldsymbol{\Psi}_0)$.

The approximate marginal posterior distributions of the model parameters, using (3) and as explained earlier for the standard model, are obtained as follows:

- i. $q(\boldsymbol{\alpha}) \stackrel{d}{=} \mathcal{N}(\hat{\boldsymbol{\alpha}}, \hat{\mathbf{A}})$ where

$$\hat{\boldsymbol{\alpha}} = \langle \sigma^{-2} \rangle \hat{\mathbf{A}} \sum_{i=1}^n \mathbf{X}'_i (\mathbf{y}_i - \mathbf{Z}_i \langle \boldsymbol{\Lambda} \rangle \langle \mathbf{b}_i \rangle) \quad (8)$$

$$\hat{\mathbf{A}} = \langle \sigma^{-2} \rangle^{-1} \left(\sum_{i=1}^n \mathbf{X}'_i \mathbf{X}_i + \langle \sigma^{-2} \rangle^{-1} \langle \boldsymbol{\Lambda}^{-1} \rangle \right)^{-1} \quad (9)$$

ii. $q(\sigma^2) \stackrel{d}{=} \mathcal{IG}(\hat{c}, \hat{d})$ where

$$\hat{c} = \sum_{i=1}^n n_i/2 + c_0 \quad (10)$$

$$\begin{aligned} \hat{d} = \frac{1}{2} \sum_{i=1}^n & \left(\mathbf{y}'_i \mathbf{y}_i - 2\langle \boldsymbol{\alpha}' \mathbf{X}'_i \mathbf{y}_i - 2\langle \mathbf{b}_i \rangle' \langle \boldsymbol{\Lambda} \rangle \mathbf{Z}'_i \mathbf{y}_i + \sum_{k=1}^p \mathbf{x}'_{ik} \langle \boldsymbol{\alpha} \boldsymbol{\alpha}' \rangle \mathbf{x}_{ik} \right. \\ & \left. + \sum_{k=1}^p \mathbf{z}'_{ik} \langle \boldsymbol{\lambda} \boldsymbol{\lambda}' \rangle \bullet \langle \mathbf{b}_i \mathbf{b}'_i \rangle \mathbf{z}_{ik} + 2\langle \boldsymbol{\alpha}' \mathbf{X}'_i \mathbf{Z}'_i \langle \boldsymbol{\Lambda} \rangle \langle \mathbf{b}_i \rangle \right) + d_0 \end{aligned} \quad (11)$$

iii. $q(\mathbf{b}_i) \stackrel{d}{=} \mathcal{N}(\hat{\mathbf{b}}_i, \hat{\mathbf{B}}_i)$ where

$$\hat{\mathbf{b}}_i = \langle \sigma^{-2} \rangle \hat{\mathbf{B}}_i \langle \boldsymbol{\Lambda} \rangle \mathbf{Z}'_i (\mathbf{y}_i - \mathbf{X}_i \langle \boldsymbol{\alpha} \rangle) \quad (12)$$

$$\hat{\mathbf{B}}_i = \langle \sigma^{-2} \rangle^{-1} \left(\langle \boldsymbol{\lambda} \boldsymbol{\lambda}' \rangle \bullet \mathbf{Z}'_i \mathbf{Z}_i + \langle \sigma^{-2} \rangle^{-1} \langle \mathbf{B}^{-1} \rangle \right)^{-1} \quad (13)$$

iv. $q(\boldsymbol{\lambda}) \stackrel{d}{=} \mathcal{N}(\hat{\boldsymbol{\lambda}}, \hat{\mathbf{V}})$ where

$$\hat{\boldsymbol{\lambda}}_i = \langle \sigma^{-2} \rangle \hat{\mathbf{V}} \sum_{i=1}^n \text{diag}(\langle \mathbf{b}_i \rangle) \mathbf{Z}'_i (\mathbf{y}_i - \mathbf{X}_i \langle \boldsymbol{\alpha} \rangle) \quad (14)$$

$$\hat{\mathbf{V}} = \langle \sigma^{-2} \rangle^{-1} \left(\sum_{i=1}^n \langle \mathbf{b}_i \mathbf{b}'_i \rangle \bullet \mathbf{Z}'_i \mathbf{Z}_i + \langle \sigma^{-2} \rangle^{-1} \langle \mathbf{V}^{-1} \rangle \right)^{-1} \quad (15)$$

v. $q(\mathbf{B}) \stackrel{d}{=} \mathcal{IW}(\hat{n}, \hat{\boldsymbol{\Psi}})$ where

$$\hat{n} = n + n_0 \quad (16)$$

$$\hat{\boldsymbol{\Psi}} = \sum_{i=1}^n \langle \mathbf{b}_i \mathbf{b}'_i \rangle + \boldsymbol{\Psi}_0 \quad (17)$$

vi. $q(\alpha_k^{-1}) \stackrel{d}{=} \mathcal{G}(\hat{\eta}, \hat{\zeta}_k)$ where

$$\hat{\eta} = 1/2 + \eta_0 \quad (18)$$

$$\hat{\zeta}_k = \langle \alpha_k^2 \rangle / 2 + \zeta_0 \quad (19)$$

vii. $q(v_k^{-1}) \stackrel{d}{=} \mathcal{G}(\eta^*, \zeta_k^*)$ where

$$\eta^* = 1/2 + \eta_0 \quad (20)$$

$$\zeta_k^* = \langle \lambda_k^2 \rangle / 2 + \zeta_0 \quad (21)$$

Here $\boldsymbol{\lambda} = \text{diag}(\boldsymbol{\Lambda})$, $\mathbf{A} = \text{diag}(a_k : k = 1, \dots, q)$, $\mathbf{V} = \text{diag}(v_k : k = 1, \dots, q)$, (\bullet) denotes the Hadamard product and $\text{diag}(\cdot)$, depending on its argument, either builds a vector from the diagonal elements of a matrix or builds a diagonal matrix using the components of a vector as the diagonal elements of that matrix.

The required moments are $\langle \boldsymbol{\alpha} \rangle = \hat{\boldsymbol{\alpha}}$, $\langle \boldsymbol{\alpha} \boldsymbol{\alpha}' \rangle = \hat{\mathbf{A}} + \hat{\boldsymbol{\alpha}} \hat{\boldsymbol{\alpha}}'$, $\langle \mathbf{b}_i \rangle = \hat{\mathbf{b}}_i$, $\langle \mathbf{b}_i \mathbf{b}_i' \rangle = \hat{\mathbf{B}}_i + \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i'$, $\langle \sigma^{-2} \rangle = \hat{c}/\hat{d}$, $\langle \boldsymbol{\Lambda} \rangle = \text{diag}(\hat{\boldsymbol{\lambda}})$, $\langle \boldsymbol{\lambda} \boldsymbol{\lambda}' \rangle = \hat{\mathbf{V}} + \hat{\boldsymbol{\lambda}} \hat{\boldsymbol{\lambda}}'$, $\langle a_k^{-1} \rangle = \hat{\eta}/\hat{\zeta}_k$, $\langle v_k^{-1} \rangle = \eta^*/\zeta_k^*$, $\langle \mathbf{B}^{-1} \rangle = \hat{n} \hat{\boldsymbol{\Phi}}^{-1}$.

The expression for the lower-bound, \mathcal{L} , is given by

$$\begin{aligned} \mathcal{L} &= \langle \log p(\mathbf{y} | \boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\lambda}, \sigma^2) \rangle + \langle \log p(\boldsymbol{\alpha} | \mathbf{a}^{-1}) \rangle + \langle \log p(\mathbf{a}^{-1}) \rangle + \langle \log p(\mathbf{b} | \mathbf{B}) \rangle \\ &\quad + \langle \log p(\mathbf{B}) \rangle + \langle \log p(\boldsymbol{\lambda} | \mathbf{v}^{-1}) \rangle + \langle \log p(\mathbf{v}^{-1}) \rangle - \langle \log q(\boldsymbol{\alpha}) \rangle - \langle \log q(\mathbf{a}^{-1}) \rangle \\ &\quad - \langle \log q(\mathbf{b}) \rangle - \langle \log q(\mathbf{B}) \rangle - \langle \log q(\boldsymbol{\lambda}) \rangle - \langle \log q(\mathbf{v}^{-1}) \rangle \\ &= \frac{1}{2} \left\{ - \sum_{i=1}^n n_i \log(2\pi) + q(n+1) + p + \log |\mathbb{V}(\boldsymbol{\alpha})| + \log |\mathbb{V}(\boldsymbol{\lambda})| \right. \\ &\quad \left. + \sum_{i=1}^n \log |\mathbb{V}(\mathbf{b}_i)| + q(\hat{n} - n_0) \log 2 + \log \frac{|\boldsymbol{\Psi}_0|^{n_0}}{|\hat{\boldsymbol{\Psi}}|^{\hat{n}}} \right\} + \log \frac{\Gamma_q(\hat{n}/2)}{\Gamma_q(n_0/2)} + \log \frac{d_0^{c_0}}{\hat{d}^{\hat{c}}} \\ &\quad + \log \frac{\Gamma(\hat{c})}{\Gamma(c_0)} + \sum_{j=1}^p \log \frac{\zeta_0^{\eta_0}}{\hat{\zeta}_j^{\hat{\eta}}} + \sum_{j=1}^q \log \frac{\zeta_0^{\eta_0}}{\zeta_j^{*\eta^*}} + p \log \frac{\Gamma(\hat{\eta})}{\Gamma(\eta_0)} + q \log \frac{\Gamma(\eta^*)}{\Gamma(\eta_0)}. \quad (22) \end{aligned}$$

4. Simulations

We now demonstrate the gain and advantages through the model explained in Section 3.2. Here we will study how closely we can estimate the fixed effect and the random effects covariance.

We specify two levels of subject size, $n = \{400, 2000\}$, three levels of number of potential covariates, $p = \{4, 20, 60\}$, $q = p$, and three levels of underlying sparsity corresponding respectively to the number of potential covariates, $p' = \{.75p, .50p, .25p\}$, $q' = p'$, where p' and q' denote the number of active covariates in the underlying model. We generate $n_i = 8$, $i = 1, \dots, n$ observations per subject as before and set $\alpha_{1:p'} = \mathbf{1}$ and $\alpha_{(p'+1):p} = \mathbf{0}$. The rest

is randomized in the following fashion for each of the 100 data sets generated: $x_{ij1} = 1$, $\mathbf{x}_{ij(2:p)} \sim \mathcal{N}_{p-1}(\mathbf{0}, \mathbf{C})$ and $\mathbf{C} \sim \mathcal{W}(p-1, \mathbf{I}_{p-1})$; $\mathbf{z}_{ij} = \mathbf{x}_{ij}$; $\sigma^2 \sim \mathcal{U}(1, 3)$; $\mathbf{D}_{1:q' \times 1:q'} \sim \mathcal{W}(q', \mathbf{I}_{q'})$ and the rest of the entries along the dimensions $(q'+1) : q$ are 0.

We run the variational procedure both for the standard and shrinkage models. For the standard model, the priors are specified as $\alpha \sim N(\mathbf{0}, \mathbf{A}_0)$, $\sigma^{-2} \sim \mathcal{G}(c_0, d_0)$ and $\mathbf{D} \sim \mathcal{IW}(n_0, \Psi_0)$ where $\alpha_0 = 0$, $\mathbf{A}_0 = 1000$ (Chen and Dunson, 2003), $c_0 = 0.1$, $d_0 = 0.001$ (Smith and Kohn, 2002), $n_0 = q$ and $\Phi_0 = \mathbf{I}$ to reflect our vague prior information on α , σ^{-2} and \mathbf{D} . All these priors are proper yet specify very vague information about the parameters relatively to the likelihoods that are observed in this simulation. For the shrinkage model, we choose $c_0 = \eta_0 = 0.1$, $d_0 = \zeta_0 = 0.001$, $n_0 = q$ and $\Phi_0 = \mathbf{I}$ to express our vague information on σ^{-2} , a_k^{-1} , v_k^{-1} and \mathbf{B} . It is important that we refrain from using improper priors on the higher level parameters, i.e. a_k^{-1} , v_k^{-1} , for meaningful marginal likelihoods as the limiting cases of these conjugate priors will lead to the impropriety of the posterior distribution and consequently the decomposition in (1) will lose its meaning.

The boxplots in Figure 1 (a) and (b) give the quadratic losses in the estimation of α and \mathbf{D} respectively arising from the standard and the shrinkage models. As the dimension of the problem increases and the underlying model becomes sparser, the advantage of the shrinkage model is highly pronounced. Figure 2 demonstrates the shrinkage toward 0 on the diagonals of the random effects covariance matrix which are 0 in the underlying model. As expected, the shrinkage model gives much better estimates for 0-diagonals.

5. Real Data Application

Here we apply the proposed method to US Collaborative Perinatal Project (CPP) data on maternal pregnancy smoking in relation to child growth (Chen et al., 2005). (Chen et al., 2005) examine the relationship between maternal smoking habits during pregnancy and childhood obesity within $n = 34866$ children in the CPP using generalized estimating equations (GEE) (Liang and Zeger, 1986). The size of the data hinders a random effects analysis (Pennell and Dunson, 2007).

Having removed the missing observations we were left with 28211 subjects and 115811 observations. We set aside 211 observations across the subjects as a hold-out sample to test the performance of our procedure which leaves us with 115600 observations to train our model with. Our design matrix, $\mathbf{X} = \mathbf{Z}$, (with a column of 1s for the intercept term) has 72 columns. Each column of $\mathbf{X} = \mathbf{Z}$ is

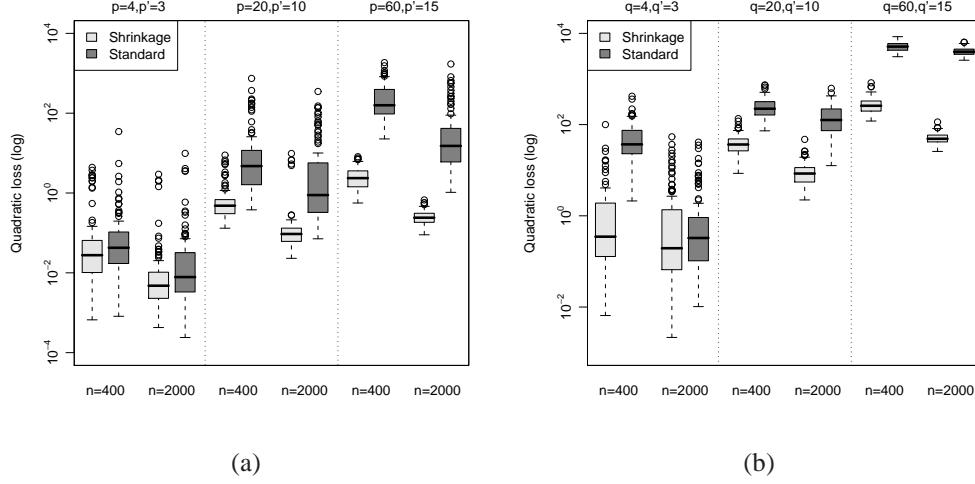


Figure 1: Quadratic loss for (a) α and (b) D . The vertical axis is given in log-scale.

scaled to have unit length. The response was the weight of the children in kilograms. A detailed description and analysis of the data set will not be provided as the main purpose here is to demonstrate the applicability of the proposed method on such data sets and to observe the shrinkage effect.

We apply our shrinkage model to the data. Figure 3 (a) and (b) give the 99% credible sets for the fixed effect coefficients and for the diagonals of the random effects covariance matrix respectively. We can see for both fixed effects coefficients and the diagonals of the random effects covariance, except for a few dimensions, most of the credible sets are concentrated around 0. Figure 3 (c) also gives the point estimates and 99% credible sets for the hold-out sample. Here the R^2 on the test set was found to be 94.8%.

The computational advantage of the procedure is undeniable. For the shrinkage model, the algorithm was implemented in MATLAB on a computer with a 2.8 GHz processor and 12 GB RAM. Figure 3(d) tracks the lower-bound and the relative error between two subsequent lower-bound values, $\psi = |\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}|/|\mathcal{L}^{(t)}|$, for convergence where $\mathcal{L}^{(t)}$ denotes the lower-bound evaluated at iteration t . The preset value of $\psi = 10^{-6}$ is reached after 2485 iterations which takes 208332 seconds. It should be noted that the computational intensity for one iteration is almost identical to a Gibbs sampling scenario, which suggests, if Gibbs sampling procedure were to be used, only 2485 samples would have been drawn. Considering the burn-in period required for convergence and the thinning of the chain to obtain

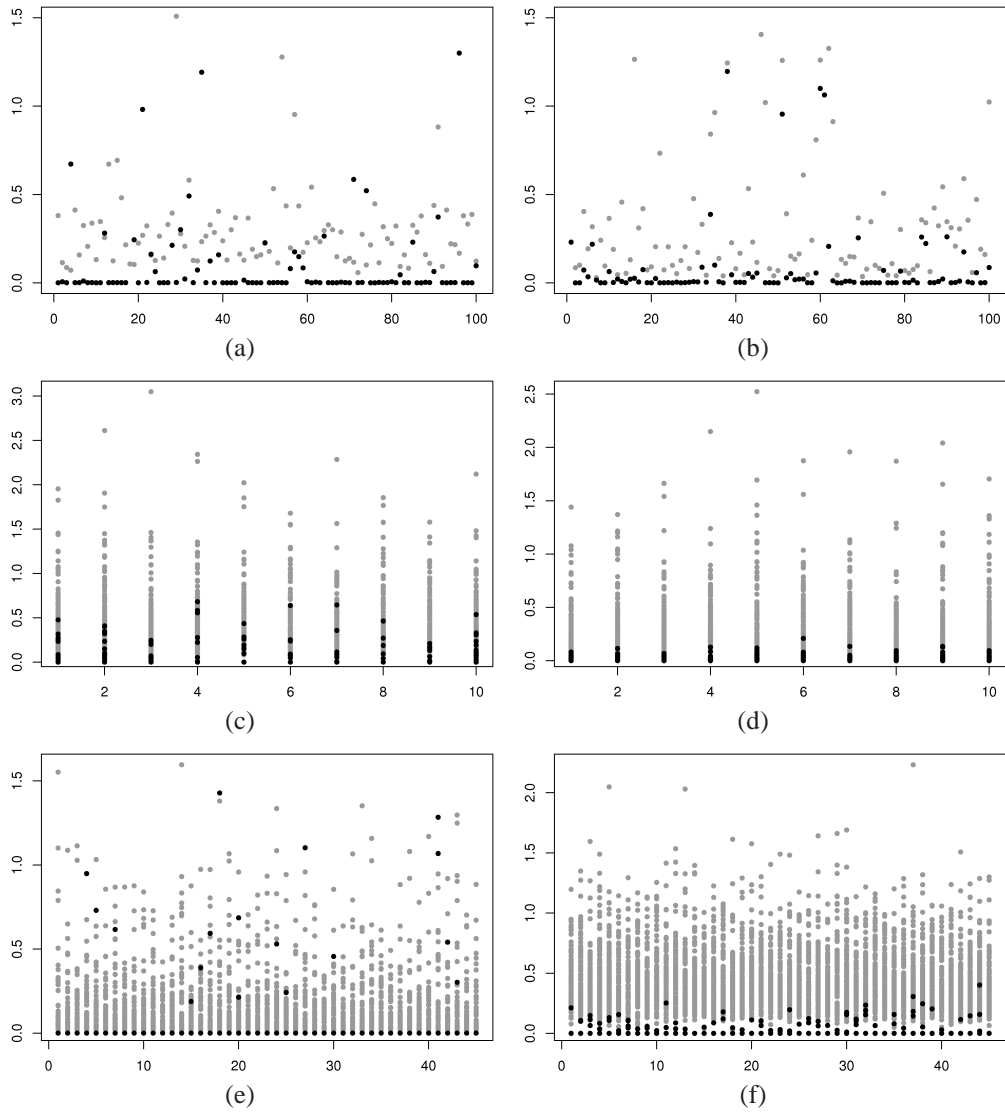


Figure 2: Estimates of 0 diagonals of \mathbf{D} (black: shrinkage model, grey: standard model). The left and right columns are respectively for $n = \{400, 2000\}$ and the rows from top to bottom respectively are for $(q, q') = \{(4, 3), (20, 10), (60, 45)\}$. Since in the first row, there is only one 0-diagonal, 100 cases are plotted along the horizontal axis while for the remainder they are plotted along the vertical axis.

less correlated draws, this number is far from sufficient. Thus, with data sets this large or larger, MCMC is not a computationally feasible option.

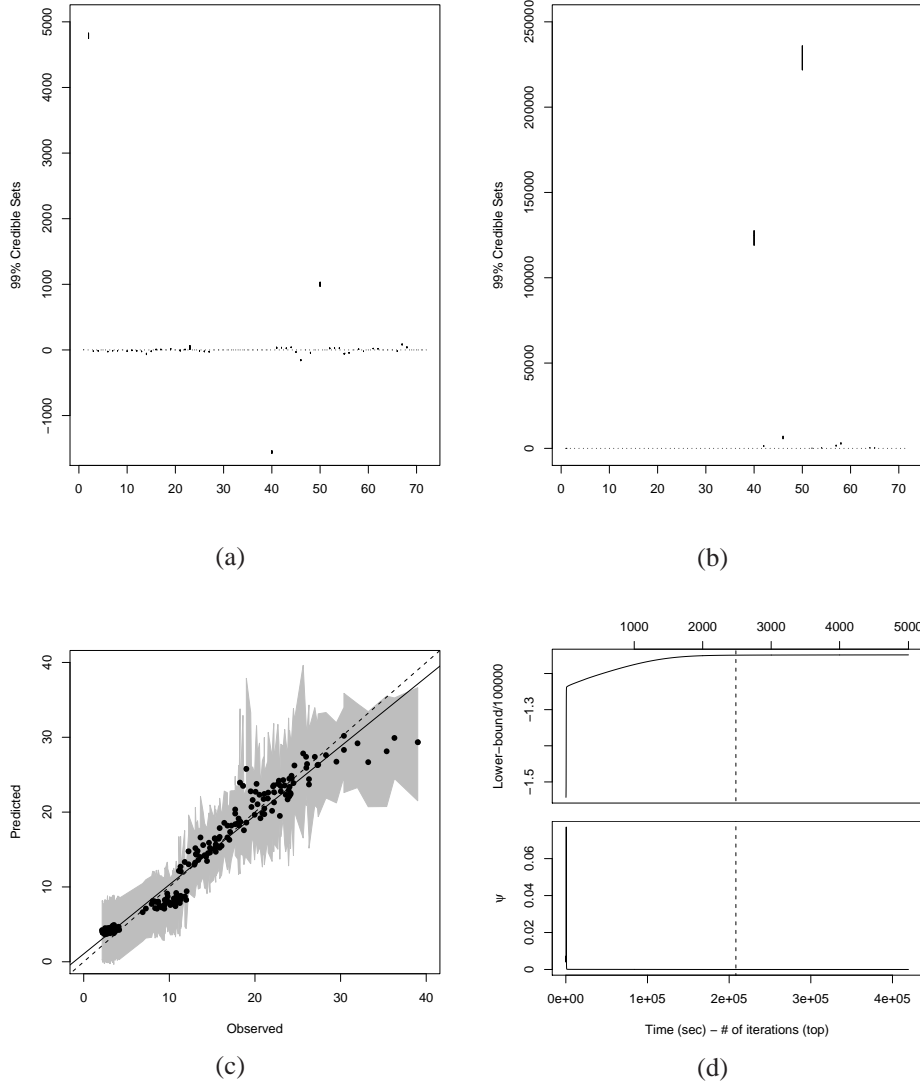


Figure 3: Vertical lines represent 99% credible sets for (a) the fixed effect coefficients for 72 predictors used in the model and (b) the diagonal elements of the random effects covariance matrix. (c) gives the predicted vs. observed plot where black circles represent the point estimates for the shrinkage model, dashed line is the 45° line and the solid line is the linear fit between the predicted and observed values. Shaded are gives a 99% credible region. (d) tracks the lower-bound (upper) and ψ for convergence (lower) over time. Vertical dashed line marks the iteration/time the pre-specified convergence criterion is reached ($\psi = 10^{-6}$).

6. Conclusion

Here we provided a fast approximate solution to fully Bayes inference to be used in the analysis of *large* longitudinal data sets. The proposed parameterization also allows for identifying the predictors that contribute as fixed and/or random effects. Although this parameterization leads to an unidentifiable likelihood, and would also cause the so-called label-switching problem with the application of Gibbs sampling, the variational approach allows us to converge to one of many solutions which lead to identical inferences. The utility of the new parameterization is justified through a simulation study. The application to a large epidemiological data set also demonstrates computational advantages obtained through the proposed method over conventional sampling techniques.

References

- Bishop, C. M. and Tipping, M. E. “Variational Relevance Vector Machines.” In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, 46–53. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. (2000).
- Chen, A., Pennell, M., Klebanoff, M. A., Rogan, W. J., and Longnecker, M. P. “Maternal smoking during pregnancy in relation to child overweight: follow-up to age 8 years.” *International Journal of Epidemiology* (2005).
- Chen, Z. and Dunson, D. B. “Random Effects Selection in Linear Mixed Models.” *Biometrics*, 59(4):762–769 (2003).
- Figueiredo, M. A. T. “Adaptive sparseness for supervised learning.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159 (2003).
- Foster, Scott, D., Verbyla, Arunas, P., Pitchford, and Wayne, S. *Journal of Agricultural, Biological and Environmental Statistics*, 12(2):300–314 (2007).
- Frühwirth-Schnatter, S. and Tüchler, R. “Bayesian parsimonious covariance estimation for hierarchical linear mixed models.” *Statistics and Computing*, 18(1):1–13 (2008).
- George, E. and McCulloch, R. “Approaches for Bayesian variable selection.” *Statistica Sinica*, 7:339–373 (1997).

- Huang, Z. and Gelman, A. “Sampling for Bayesian computation with large datasets.” *Technical Report, Department of Statistics, Columbia University*, 7:339–373 (2008).
- Jaakkola, T. S. and Jordan, M. I. “Bayesian parameter estimation via variational methods.” *Statistics and Computing*, 10(1):25–37 (2000).
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. “An introduction to variational methods for graphical models.” 105–161 (1999).
- Kinney, S. K. and Dunson, D. B. “Fixed and Random Effects Selection in Linear and Logistic Models.” *Biometrics*, 63(3):690–698 (2007).
- Laird, N. M. and Ware, J. H. “Random-Effects Models for Longitudinal Data.” *Biometrics*, 38(4):963–974 (1982).
- Liang, K.-Y. and Zeger, S. L. “Longitudinal data analysis using generalized linear models.” *Biometrika*, 73(1):13–22 (1986).
- Park, T. and Casella, G. “The Bayesian Lasso.” *Journal of the American Statistical Association*, 103:681–686(6) (2008).
- Pennell, M. L. and Dunson, D. B. “Fitting semiparametric random effects models to large data sets.” *Biostat*, 8(4):821–834 (2007).
- Smith, M. and Kohn, R. “Parsimonious Covariance Matrix Estimation for Longitudinal Data.” *Journal of the American Statistical Association*, 97:1141–1153 (2002).
- Tibshirani, R. “Regression Shrinkage and Selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288 (1996).
- Tierney, L. and Kadane, J. B. “Accurate Approximations for Posterior Moments and Marginal Densities.” *Journal of the American Statistical Association*, 81(393):82–86 (1986).
- Tipping, M. E. “Sparse Bayesian Learning and the Relevance Vector Machine.” *Journal of Machine Learning Research*, 1 (2001).
- West, M. “On Scale Mixtures of Normal Distributions.” *Biometrika*, 74(3):646–648 (1987).