Assessing the performance of variational methods for mixed logistic regression models*

Frank Rijmen[1,2] and Jiří Vomlel[3]

[1] Department of Psychology
K.U.Leuven
Leuven, Belgium

[2] Department of Clinical Epidemiology and Biostatistics
VU University Medical Center,
Amsterdam, The Netherlands.
F.Rijmen@wumc.nl

[3] Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4
182 08, Prague, Czech Republic
Tel.+420-266-052-395
Fax +420-286-581-419
vomlel@utia.cas.cz

Abstract. We present a variational estimation method for the mixed logistic regression model. The method is based on a lower bound approximation of the logistic function (Jaakkola and Jordan, 2000). Based on the approximation, an EM algorithm can be derived that results in a considerable simplification of the maximization problem in that it does not require the numerical evaluation of integrals over the random effects. We assess the performance of the variational method for the mixed logistic regression model in a simulation study, and compare it to Laplace's method. The results indicate that the variational method is a viable choice for estimating the fixed effects of the mixed logistic regression model under the condition that the number of outcomes within each cluster is sufficiently high.

Keywords. Mixed models, logistic regression, variational methods, lower bound approximation.
MSC. 62J12, 62P15, 62E17

1 Introduction

Mixed models are one of the standard tools for the analysis of clustered data. Clustered data arise for example in the context of longitudinal studies, where a sample of clusters is repeatedly assessed; or in educational settings where pupils are clustered within schools. The outcomes stemming from the same cluster tend to be more homogeneous than outcomes stemming from different clusters, and, by consequence, the outcomes within a cluster are likely to be correlated. These dependencies are taken into account by mixed models through the incorporation of so called random effects. These are cluster-specific parameters that are considered to be random variables as well, with a distribution that is defined over the population of clusters. The cluster-specific parameters are assumed to explain away all dependencies that are due to inter-cluster variability.

* This is a preprint of an article submitted for consideration in the STATISTICAL COMPUTATION AND SIMULATION . 2007 ©Taylor & Francis; STATISTICAL COMPUTATION AND SIMULATION is available online at: http://journalsonline.tandf.co.uk/

Let y_{ij} be the j-th observed outcome of cluster i, $i = 1, \ldots, N$, $j = 1, \ldots, J_i$, and $\boldsymbol{b}_i$ a vector of parameters that are specific for cluster i (random effects). In its general form, a mixed model can be defined as follows:

$$y_{ij} \text{ are independent realizations of } f(y_{ij} \mid \boldsymbol{b}_i), \text{ conditional on } \boldsymbol{b}_i \tag{1}$$

$$\boldsymbol{b}_i \text{ are independent realizations of } k(\boldsymbol{b}_i), \tag{2}$$

By the assumption of conditional independence between the outcomes within a cluster,

$$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \prod_{j=1}^{J_i} f(y_{ij} \mid \boldsymbol{b}_i) \ .$$

Most commonly, a (multivariate) normal distribution is assumed for the random effects,

$$k(\boldsymbol{b}_i) = \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma}) \ .$$

The choice for the conditional distribution of the outcomes determines to which class of mixed models the model belongs:

- In the linear mixed model (Verbeke and Molenberghs, 2000; Diggle et al., 2002)

  $$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{\mu}_i, \boldsymbol{R}) \ \text{ with } \boldsymbol{\mu}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,$$

  where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ denote the predictor matrices for the fixed and random effects, respectively, and $\boldsymbol{\beta}$ denotes the vector of parameters that are common to all clusters (fixed effects)
- In the nonlinear mixed model (Davidian and Giltinan, 1995)

  $$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{R}) \ \text{ with } \boldsymbol{\mu}_i = m(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i),$$

  where m is a nonlinear function.
- In the generalized linear mixed model (McCulloch and Searle, 2001),

  $$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i) = \prod_{j=1}^{J_i} f(y_{ij} \mid \boldsymbol{b}_i) \ ,$$

  with $f(y_{ij} \mid \boldsymbol{b}_i)$ being a member of the exponential family. The conditional mean μ_{ij} of $f(y_{ij} \mid \boldsymbol{b}_i)$ is related to the predictors via the link function g:

  $$g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i \ .$$

Note that, if $\boldsymbol{R}$ is not the identity matrix, then for linear and nonlinear mixed models, we obtain a more general mixed model than one presented in equations (1) and (2) in the sense that conditional independence between the outcomes of the same cluster is no longer imposed. This allows for the modeling of additional sources of dependencies (e.g. serial correlation).

Maximum likelihood estimates for the parameters are obtained by maximizing the log-likelihood

$$\ell(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \log \left(\int_{-\infty}^{+\infty} p(\boldsymbol{y}_i \mid \boldsymbol{b}_i) \cdot k(\boldsymbol{b}_i) \ d\boldsymbol{b}_i \right) \ , \tag{3}$$

where $\boldsymbol{\alpha}$ consists of the parameter vector $\boldsymbol{\varphi}$ that characterizes the distribution of the random effects $k(\boldsymbol{b}_i)$, $\boldsymbol{\beta}$, and the remaining parameters that characterize $p(\boldsymbol{y}_i \mid \boldsymbol{b}_i)$ (e.g., the parameters that determine $\boldsymbol{R}$ in the (non)linear mixed model).

For the linear mixed model with a normal distribution for the random effects and for some generalized linear mixed models, the integral in equation (3) has a closed form solution. However, for most commonly used generalized linear and nonlinear mixed model, the integral is intractable.

A variety of approximations to the integrals appearing in the log-likelihood have been proposed. They can be categorized into two categories (Tuerlinckx et al., 2006): *approximations to the integral* on the one hand, and *approximations to the integrand* on the other hand. The gaussian quadrature methods (adaptive or nonadaptive) are methods in which the integral is approximated by a finite sum. Examples of methods in which the integrand is approximated are Laplace's method (Tierny and Kadane, 1986), Marginal Quasi-Likelihood (MQL) (Goldstein, 1991), Penalized Quasi-Likelihood (PQL) (Breslow and Clayton, 1993) and extensions of these methods (Raudenbush et al., 2000; Goldstein and Rasbash, 1996; Rodriguez and Goldman, 1995). For a review, see Tuerlinckx et al. (2006).

In this paper, we present and evaluate a method that belongs to a third category of approximations: the category of variational methods. The distinguishing feature of variational methods is the introduction of so called *variational parameters* to obtain a lower bound approximation to the log-likelihood.

2 A variational method for the mixed logistic regression model

Variational techniques encompass a variety of approximation techniques. The common denominator is to simplify a complex optimization problem by the introduction of additional, variational, parameters. For a fixed set of values for the variational parameters, the transformed problem has a simpler (e.g. closed-form) solution, providing an approximate solution to the original problem. The variational parameters are optimized in a separate step. Estimation is performed by alternating these two steps in order to obtain a sequence of approximations of increasing accuracy (Jaakkola and Jordan, 2000). Examples of the use of variational lower bounds to the log-likelihood function can be found in the context of missing data for Markovian models (Hall et al., 2002), and in the context of graphical models (Jordan et al., 1999; Saul et al., 1996).

Jaakkola and Jordan (2000) proposed a variational method to obtain a closed form approximation to the posterior distribution of the parameters of the logistic regression model formulated in a Bayesian framework. Their method is based on a lower bound variational approximation of the logistic function

$$h(x) = \frac{\exp(x)}{1+\exp(x)} \quad = \quad (1+\exp(-x))^{-1}$$

proposed by Jaakkola and Jordan (2000), see also Tipping (1998).

Using the same lower bound variational approximation to the logistic function, we will present a method for obtaining maximum likelihood estimates of the parameters of a mixed logistic regression model with a normal distribution for the random effects. The method is a natural extension of the work of Jaakkola and Jordan (2000).

The mixed logistic regression model is a generalized linear mixed model with a binomial distribution as the choice for the exponential family distribution, and the logit function as a link function. Without loss of generality, we only consider the case of binary data. The log-likelihood function is then

$$\begin{aligned} \ell(\boldsymbol{\alpha}) &= \sum_{i=1}^{N} \log \int_{-\infty}^{+\infty} \mathcal{N}(\boldsymbol{b}_i; \mathbf{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{J_i} h\left((2y_{ij}-1)(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i)\right) d\boldsymbol{b}_i \\ &= \sum_{i=1}^{N} \log \int_{-\infty}^{+\infty} \mathcal{N}(\boldsymbol{b}_i; \mathbf{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{J_i} h\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right) d\boldsymbol{b}_i \quad , \end{aligned} \tag{4}$$

where we simplified notation by defining the function

$$t(y, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{b}) = (2y-1)(\boldsymbol{x}'\boldsymbol{\beta} + \boldsymbol{z}'\boldsymbol{b}) \ ,$$

which is linear in $\boldsymbol{b}$. When it is clear from the context we omit other arguments of the function and write $t(\boldsymbol{b})$.

The main distinction between the mixed logistic regression model and the "Bayesian" logistic regression model of Jaakkola and Jordan (2000) are the clustered data structure and the presence of both fixed and random effects in the former, whereas in a Bayesian framework, all parameters are considered to be random variables. Jaakkola and Jordan (2000) discuss clustered data to some extent in the context of binary factor analysis, see also Tipping (1998).

Because the mixed logistic regression model is related but not equivalent to the models discussed by Jaakkola and Jordan (2000), the derivations for the variational method are presented to some detail in the following.

Variational lower bound to the log-likelihood

The variational approximation to the integral in equation (4) starts by constructing a variational lower bound to the logistic function. First, the log logistic function is symmetrized

$$\log h(x) = \frac{x}{2} - \log\left(\exp(\frac{x}{2}) + \exp(-\frac{x}{2})\right) \ . \tag{5}$$

The second term $f(x) = -\log\left(\exp(\frac{x}{2}) + \exp(-\frac{x}{2})\right)$ is convex in x^2, so that the tangent to its surface formed by a first order Taylor expansion in x^2 around a point ξ is a lower bound

$$\begin{aligned} f(x) &\geq f(\xi) + \frac{\partial f(x)}{\partial (x^2)}\Big|_{x=\xi}(x^2 - \xi^2) \\ &= f(\xi) - \lambda(\xi)(x^2 - \xi^2) \ , \end{aligned}$$

where

$$\lambda(\xi) = -\frac{\partial f(x)}{\partial (x^2)}\Big|_{x=\xi} \quad = \quad \frac{1}{4\xi}\tanh(\frac{\xi}{2}) \ . \tag{6}$$

It implies that

$$\begin{aligned} h(x) &\geq \exp\left(\frac{x}{2} + f(\xi) - \lambda(\xi)(x^2 - \xi^2)\right) \\ &= \exp\left(\frac{\xi}{2} + f(\xi)\right)\exp\left(\frac{x}{2} - \frac{\xi}{2} - \lambda(\xi)(\xi^2 - h^2)\right) \\ &= h(\xi)\exp\left(\frac{1}{2}(x-\xi) - \lambda(\xi)(x^2 - \xi^2)\right) \ . \end{aligned}$$

In Figure 1 we present logistic function together with its variational lower bounds for five different values of variational parameter ξ.

We will use the lower bound in the context of the mixed logistic regression model and approximate $h(t(\boldsymbol{b}))$ by its variational lower bound

$$\underline{h}(t(\boldsymbol{b})) = h(\xi)\exp\left(\frac{1}{2}\cdot(t(\boldsymbol{b}) - \xi) - \lambda(\xi)\cdot(t(\boldsymbol{b})^2 - \xi^2)\right) \ .$$

Allowing a different value ξ_{ij} of the variational parameter for each outcome j within in each cluster i, a lower bound for the log-likelihood is obtained by

$$\ell(\boldsymbol{\alpha}) \ \geq \ \underline{\ell}(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\log\int_{-\infty}^{+\infty}\mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma})\prod_{j=1}^{J_i}\underline{h}\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right)d\boldsymbol{b}_i \ . \tag{7}$$

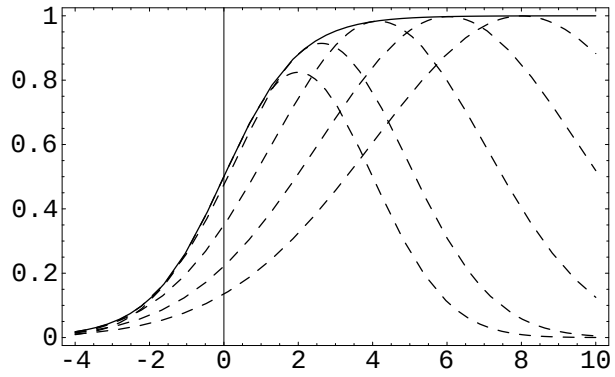


Fig. 1. Logistic function (solid line) and its variational lower bounds for $\xi = 0, +2, +4, +6, +8$ (dashed lines)

Normality of the approximated posterior distribution of the random effects

It is easily verified that the exponent of the variational lower bound is quadratic in $\boldsymbol{b}$

$$\begin{aligned}\underline{h}(t(\boldsymbol{b})) &\propto \exp\left(\frac{1}{2}\cdot t(\boldsymbol{b}) - \lambda(\xi)\cdot t(\boldsymbol{b})^2\right)\\ &\propto \exp\left((y-\frac{1}{2})\boldsymbol{z}'\boldsymbol{b} - \lambda(\xi)(2\boldsymbol{x}'\boldsymbol{\beta}\boldsymbol{z}'\boldsymbol{b} + \boldsymbol{b}'\boldsymbol{z}\boldsymbol{z}'\boldsymbol{b})\right)\\ &= \exp\left(\left((y-\frac{1}{2})\boldsymbol{z}' - \lambda(\xi)2\boldsymbol{x}'\boldsymbol{\beta}\boldsymbol{z}'\right)\boldsymbol{b} - \lambda(\xi)\boldsymbol{b}'\boldsymbol{z}\boldsymbol{z}'\boldsymbol{b})\right),\end{aligned}$$

Hence, as a function of $\boldsymbol{b}$, $\underline{h}(t(\boldsymbol{b}))$ is proportional to a normal distribution with mean $\boldsymbol{\nu}$ and covariance matrix Ψ, where

$$\begin{aligned}\Psi^{-1} &= 2\lambda(\xi)\boldsymbol{z}\boldsymbol{z}'\\ \boldsymbol{\nu} &= \Psi\left(y - \frac{1}{2} - 2\lambda(\xi)\boldsymbol{x}'\boldsymbol{\beta}\right)\boldsymbol{z}\ .\end{aligned}$$

This property can be used to approximate the posterior distribution of the random effects $q(\boldsymbol{b}_i \mid \boldsymbol{y}_i)$. The integrand in equation (4) is the unnormalized posterior of $\boldsymbol{b}_i$, with the integrand in equation (7) as its lower bound. As a function of $\boldsymbol{b_i}$, the latter is a product of a normal distribution and J_i factors that are proportional to a normal distribution. This implies that the lower bound of the unnormalized posterior of $\boldsymbol{b}_i$ is proportional to a normal distribution as well,

$$\mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma}) \prod_{j=1}^{J_i} \underline{h}\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right) \propto \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\ ,$$

where

$$\boldsymbol{\Sigma}_i^{-1} = \boldsymbol{\Sigma}^{-1} + 2\sum_{j=1}^{J_i}\lambda(\xi_{ij})\boldsymbol{z}_{ij}\boldsymbol{z}_{ij}' \tag{8}$$

$$\boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i \sum_{j=1}^{J_i}\left(y_{ij} - \frac{1}{2} - 2\lambda(\xi_{ij})\boldsymbol{x}_{ij}'\boldsymbol{\beta}\right)\boldsymbol{z}_{ij}\ . \tag{9}$$

Consequently, we can approximate $q(\boldsymbol{b}_i \mid \boldsymbol{y}_i)$ with $q^*(\boldsymbol{b}_i \mid \boldsymbol{y}_i) = \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Note that, by normalizing the variational approximation, it becomes a proper density and thus $\mathcal{N}(\boldsymbol{b}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is not a lower bound to the posterior distribution of the random effects $q(\boldsymbol{b}_i \mid \boldsymbol{b}_i)$ (Jaakkola and Jordan, 2000).

Alternatively, $q^*(\boldsymbol{b}_i \mid \boldsymbol{y}_i)$ is the posterior distribution of the random effects when

$$h\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right)$$

is approximated with its lower bound $\underline{h}\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right)$.

3 EM algorithm

The EM algorithm (Dempster et al., 1977) can be used to obtain parameter estimates that maximize the lower bound to the log-likelihood $\underline{\ell}(\boldsymbol{\alpha})$. Before proceeding, we take one step back and describe the "traditional" EM algorithm for obtaining maximum likelihood estimates for the mixed logistic regression model. Treating the random effects $\boldsymbol{b}_i$, $i = 1, \ldots, N$ as missing data, the complete data log-likelihood is

$$\ell_c(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \left(\log \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma}) + \sum_{j=1}^{J_i} \log h\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right) \right) \ , \tag{10}$$

In the E-step, one computes the expectation of the complete data log-likelihood given the previous parameter estimates $\hat{\boldsymbol{\alpha}}^{old}$:

$$Q(\boldsymbol{\alpha}; \hat{\boldsymbol{\alpha}}^{old}) = E_q\{\ell_c(\boldsymbol{\alpha})\} \ , \tag{11}$$

where E_q denotes the expectation with respect to the posterior distributions of the random effects $q(\boldsymbol{b}_i \mid \boldsymbol{y}_i; \hat{\boldsymbol{\alpha}}^{old})$. In the M-step, this expectation is maximized for $\boldsymbol{\alpha}$. The latter serve as new provisional parameter estimates for the next E-step. The sequence of parameter estimates converges to a stationary point of the incomplete data log-likelihood $\ell(\boldsymbol{\alpha})$.

For a mixed logistic regression model however, the E-step involves the computationally demanding numerical evaluation of integrals over the random effects $\boldsymbol{b}_i$. Therefore, we shift attention from $\ell(\boldsymbol{\alpha})$ to its lower bound $\underline{\ell}(\boldsymbol{\alpha})$, and use the EM algorithm to obtain parameter estimates that maximize the latter. The lower bound approximation of the complete data log-likelihood is

$$\underline{\ell}_c(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \left(\log \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma}) + \sum_{j=1}^{J_i} \log \underline{h}\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right) \right) \ . \tag{12}$$

E-step

The E-step involves the computation of

$$\begin{aligned} Q(\boldsymbol{\alpha}, \boldsymbol{\xi}; \hat{\boldsymbol{\alpha}}^{old}, \hat{\boldsymbol{\xi}}^{old}) \ &= \ E_{q^*}\{\underline{\ell}_c(\boldsymbol{\alpha})\} \\ = \sum_{i=1}^{N} &\left(E_{q^*}\{\log \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma})\} + \sum_{j=1}^{J_i} E_{q^*}\{\log \underline{h}\left(t(y_{ij}, \boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i)\right)\} \right) , \end{aligned} \tag{13}$$

where E_{q^*} denotes the expectation with respect to

$$q^*(\boldsymbol{b}_i \mid \boldsymbol{y}_i; \hat{\boldsymbol{\alpha}}^{old}, \hat{\boldsymbol{\xi}}^{old}) = \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{\mu}_i^{old}, \boldsymbol{\Sigma}_i^{old}) \ .$$

For the first part of the right-hand side of equation (13) we obtain

$$E_{q^*}\{\log \mathcal{N}(\boldsymbol{b}_i;\mathbf{0},\boldsymbol{\Sigma})\} = -N\log(2\pi)^{D/2} - N\log|\boldsymbol{\Sigma}|^{1/2} \\ -\frac{1}{2}\sum_{i=1}^{N}\left(\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{i}}^{\mathrm{old}}\right) + (\boldsymbol{\mu_i}^{\mathrm{old}})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i}^{\mathrm{old}}\right) \ , \tag{14}$$

where D denotes the dimensionality of $\boldsymbol{b}_i$ and we used the identity

$$E_{q^*}\{\boldsymbol{b}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{b}_i\} = \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{\mathrm{i}}^{\mathrm{old}}\right) + (\boldsymbol{\mu_i}^{\mathrm{old}})'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i}^{\mathrm{old}} \ .$$

For each term of the second part of the right hand side of equation (13) we can write

$$\begin{aligned} &E_{q^*}\{\log \underline{h}\left(t(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta},\boldsymbol{z}_{ij},\boldsymbol{b}_i)\right)\} \\ &= \log h(\xi_{ij}) + \frac{1}{2}\left(E_{q^*}\left\{t(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta},\boldsymbol{z}_{ij},\boldsymbol{b}_i)\right\} - \xi_{ij}\right) \\ &\quad -\lambda(\xi_{ij})\left(E_{q^*}\left\{t(y_{ij},\boldsymbol{x}_{ij},\boldsymbol{\beta},\boldsymbol{z}_{ij},\boldsymbol{b}_i)^2\right\} - (\xi_{ij})^2\right) \\ &= \log h(\xi_{ij}) + (y_{ij}-\frac{1}{2})(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old}) - \frac{1}{2}\xi_{ij} \\ &\quad -\lambda(\xi_{ij})\left((\boldsymbol{x}_{ij}'\boldsymbol{\beta})^2 + 2\boldsymbol{x}_{ij}'\boldsymbol{\beta}\boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old} + \boldsymbol{z}_{ij}'(\boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old})\boldsymbol{z}_{ij} - (\xi_{ij})^2\right) \ , \end{aligned} \tag{15}$$

where we used the identity $E_{q^*}\{\boldsymbol{b}_i\boldsymbol{b}_i'\} = \boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old}$.

M-step

In the M-step, the variational parameters ξ_{ij} and the model parameters $\boldsymbol{\alpha}$ are updated. The parameters that optimize $Q(\boldsymbol{\alpha},\boldsymbol{\xi};\hat{\boldsymbol{\alpha}}^{old},\hat{\boldsymbol{\xi}}^{old})$ are given by the solution to the likelihood equations.

For each of the variational parameters ξ_{ij}, $i=1,\ldots,N$, $j=1,\ldots,J_i$, the likelihood equation is

$$0 = \frac{\partial}{\partial\xi_{ij}}Q(\boldsymbol{\alpha},\boldsymbol{\xi};\hat{\boldsymbol{\alpha}}^{old},\hat{\boldsymbol{\xi}}^{old}) \tag{16}$$

$$\begin{aligned} &= \frac{\partial}{\partial\xi_{ij}}\log h(\xi_{ij}) - \frac{1}{2} + 2\lambda(\xi_{ij})\xi_{ij} \\ &\quad -\frac{\partial\lambda(\xi_{ij})}{\partial\xi_{ij}}\left((\boldsymbol{x}_{ij}'\boldsymbol{\beta})^2 + 2\boldsymbol{x}_{ij}'\boldsymbol{\beta}\boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old} + \boldsymbol{z}_{ij}'(\boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old})\boldsymbol{z}_{ij} - \xi_{ij}^2\right) \ . \end{aligned} \tag{17}$$

It follows from equations (5) and (6) that

$$\frac{\partial}{\partial\xi_{ij}}\log h(\xi_{ij}) = \frac{\partial}{\partial\xi_{ij}}\left(\frac{1}{2}\xi_{ij} + f(\xi_{ij})\right) \ = \ \frac{1}{2} + \frac{\partial f(\xi_{ij})}{\partial\xi_{ij}^2}\frac{\partial\xi_{ij}^2}{\partial\xi_{ij}} \ = \ \frac{1}{2} - 2\lambda(\xi_{ij})\xi_{ij}$$

When we substitute this into equation (17) we get

$$\begin{aligned} &\frac{\partial}{\partial\xi_{ij}}Q(\boldsymbol{\alpha},\boldsymbol{\xi};\hat{\boldsymbol{\alpha}}^{old},\hat{\boldsymbol{\xi}}^{old}) \\ &= -\frac{\partial\lambda(\xi_{ij})}{\partial\xi_{ij}}\left((\boldsymbol{x}_{ij}'\boldsymbol{\beta})^2 + 2\boldsymbol{x}_{ij}'\boldsymbol{\beta}\boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old} + \boldsymbol{z}_{ij}'\left(\boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old}\right)\boldsymbol{z}_{ij} - \xi_{ij}^2\right) \ , \end{aligned}$$

which, together with the property that $\lambda(\xi_{ij})$ is a monotonically decreasing function for $\xi_{ij} \geq 0$ (and we can always chose $\xi_{ij} \geq 0$ because $\underline{h}(t(\boldsymbol{b}))$ is a symmetric function of ξ_{ij}), gives the solution to the likelihood equation

$$\xi_{ij}^2 = (\boldsymbol{x}_{ij}'\boldsymbol{\beta})^2 + 2\boldsymbol{x}_{ij}'\boldsymbol{\beta}\boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old} + \boldsymbol{z}_{ij}'\left(\boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old}\right)\boldsymbol{z}_{ij} \tag{18}$$

The likelihood equations for the fixed parameters $\boldsymbol{\beta}$ are

$$\begin{aligned} 0 &= \frac{\partial}{\partial\boldsymbol{\beta}} Q(\boldsymbol{\alpha}, \boldsymbol{\xi}; \hat{\boldsymbol{\alpha}}^{old}, \hat{\boldsymbol{\xi}}^{old}) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J_i}(y_{ij} - \frac{1}{2})\boldsymbol{x}_{ij} + \lambda(\xi_{ij})\left(-2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}'\boldsymbol{\beta} - 2\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old}\right), \end{aligned}$$

with the solution

$$\boldsymbol{\beta} = \left(\sum_{i=1}^{N}\sum_{j=1}^{J_i} 2\lambda(\xi_{ij})\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{j=1}^{J_i}(y_{ij} - \frac{1}{2})\boldsymbol{x}_{ij} - 2\lambda(\xi_{ij})\boldsymbol{x}_{ij}\boldsymbol{z}_{ij}'\boldsymbol{\mu}_i^{old}\right)$$

Finally, the likelihood equations for the covariance parameters of of the random effects are

$$0 = \frac{\partial}{\partial\boldsymbol{\Sigma}} Q(\boldsymbol{\alpha}, \boldsymbol{\xi}; \hat{\boldsymbol{\alpha}}^{old}, \hat{\boldsymbol{\xi}}^{old}) \;=\; \sum_{i=1}^{N}\frac{\partial}{\partial\boldsymbol{\Sigma}} E_{q^*}\left\{\log\left(\mathcal{N}(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{\Sigma})\right)\right\}$$

Since

$$\frac{\partial \log|\boldsymbol{X}|}{\partial\boldsymbol{X}} \;=\; (\boldsymbol{X}^{-1})' \;,\; \frac{\partial \mathrm{tr}(\boldsymbol{X}^{-1}\boldsymbol{A})}{\partial\boldsymbol{X}} \;=\; -\boldsymbol{X}^{-1}\boldsymbol{A}'\boldsymbol{X}^{-1} \;,$$

and $\boldsymbol{\Sigma}$ is symmetric we obtain

$$0 = -N\boldsymbol{\Sigma}^{-1} + \left(\sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old}\right)\boldsymbol{\Sigma}^{-1}\right) \;.$$

The solution for $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \frac{1}{N}\left(\sum_{i=1}^{N}\boldsymbol{\mu}_i^{old}(\boldsymbol{\mu}_i^{old})' + \boldsymbol{\Sigma}_i^{old}\right)$$

Because the variational parameters ξ_{ij} appear in the solution to the likelihood equations for the fixed parameters $\boldsymbol{\beta}$, and vice versa, a conditional maximization scheme is a convenient choice for updating those two sets of parameters. That is, while updating one set of parameters, the parameters of the other set are constrained at their current values. Within each M-step, this conditional maximization scheme can be cycled through several times, but one cycle is sufficient to increase $\underline{\ell}(\boldsymbol{\alpha})$. As in the standard EM algorithm, $\underline{\ell}(\boldsymbol{\alpha}^{new}) \geq \underline{\ell}(\boldsymbol{\alpha}^{old})$ follows from Jensen's inequality.

4 Simulation study

The Rasch model

The Rasch model (Rasch, 1960) is widely used in the field of educational measurement. In this context, clusters correspond to persons, and the observations to the responses on binary items

("correct" vs. "incorrect" or "agree" vs. "disagree"). In the Rasch model, the probability of observing "1" is modeled as a logistic function of the sum of a person parameter and an item parameter

$$Pr(Y_{ij} = 1|b_i) = h(b_i + \beta_j) \ .$$

If we assume that the person parameter b_i is a random variable with a distribution over the population of persons, the Rasch model is a mixed logistic regression model with a random intercept and fixed item parameters, with

$$\begin{aligned}
\boldsymbol{x}'_{ij} = \boldsymbol{x}'_j &= (\delta_1(j), \delta_2(j), \ldots, \delta_{J_i}(j)) \ , \text{ where } \delta_k(j) = \begin{cases} 1 \text{ if } k = j \\ 0 \text{ otherwise.} \end{cases} \\
\boldsymbol{b}_i &= b_i \\
\boldsymbol{z}_{ij} = z_i &= 1 \quad \text{for } i = 1, \ldots, N \\
\boldsymbol{\Sigma} &= \sigma^2 \ .
\end{aligned}$$

Method

We simulated data for four different Rasch models, with $\sigma^2 = 1, 3$ and $J = 5, 25$. From each of these four models we generated 60 data sets: 30 data sets containing 100 "persons" and 30 datasets containing 500 "persons". Thus, we performed eight experiments:

1. $\sigma^2 = 3$, $J = 25$, and $N = 100$,
2. $\sigma^2 = 1$, $J = 25$, and $N = 100$,
3. $\sigma^2 = 3$, $J = 25$, and $N = 500$,
4. $\sigma^2 = 1$, $J = 25$, and $N = 500$,
5. $\sigma^2 = 3$, $J = 5$, and $N = 100$,
6. $\sigma^2 = 1$, $J = 5$, and $N = 100$,
7. $\sigma^2 = 3$, $J = 5$, and $N = 500$,
8. $\sigma^2 = 1$, $J = 5$, and $N = 500$.

The item parameters ranged between minus two and plus two. Furthermore, the items were placed symmetrically around zero and were placed more densely the closer they were to zero, see Figure 2.

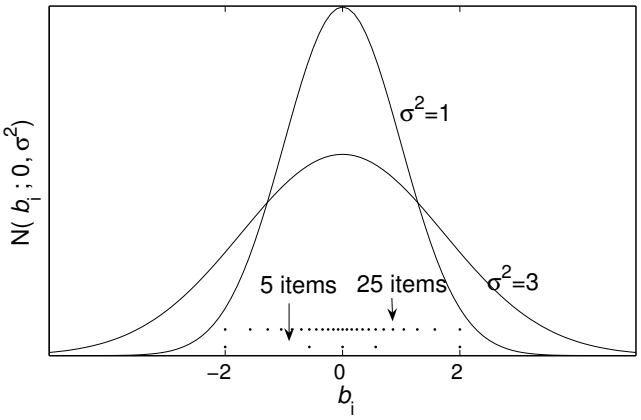
For each dataset, parameter estimates were obtained for both the variational method and Laplace's method. In Laplace's method, the log-likelihood is approximated by approximating the log integrand in equation (4) through a second-order Taylor expansion around its mode. Laplace's method is often used and will therefore serve as a benchmark for assessing the performance of the variational method.

For Laplace's method, we used the SAS NLMIXED procedure Laplace's method is obtained in SAS NLMIXED through specifying "method = gauss qpoints = 1", see the SAS user's manual (SAS Institute Inc., 2005). We implemented the variational method for the estimation of the Rasch model in Matlab.

Results

Figures 3, 4, and 5 display the boxplots of the differences between the estimated and true parameters for each of the eight conditions. The center of the boxplots are informative with respect to the bias of the estimation procedure: negative bias (underestimation of the true parameter) is indicated by a boxplot that is centered below zero, and positive bias by a boxplot that is centered above zero. The width of the boxplots are informative with respect to the

Fig. 2. Item parameters



dispersion of the estimates. In Figures 4 and 5, the boxplots for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$ are ordered in decreasing order: the leftmost boxplot corresponds to $\beta_j = 2$ and the rightmost to $\beta_j = -2$.

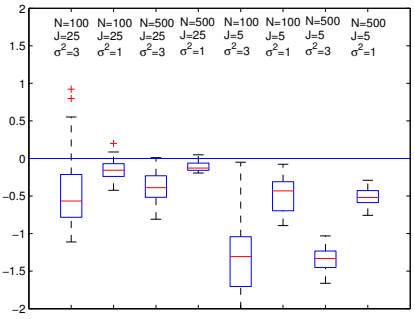
With respect to the variance parameter (figure 3), Laplace's method showed no bias when $\sigma^2 = 1$, but when $\sigma^2 = 3$, a substantial negative bias was found. The magnitude of the bias is about 50% of the true value, and is independent of sample size and the number of items. As could be expected, the dispersion of the estimates becomes smaller with increasing sample size. In addition, the estimates showed less dispersion when $J = 5$ than when $J = 25$. The variational method was found to suffer from a negative bias in all conditions. The magnitude of the bias was a function of the number of items (larger bias when $J = 5$). The absolute bias also depended on the true value of σ^2 (larger absolute bias when $\sigma^2 = 3$), but the relative bias was roughly the same (e.g. 40 to 50% for $J = 5$, for both $\sigma^2 = 1$ and $\sigma^2 = 3$). As could be expected, the dispersion of the estimates becomes smaller with increasing sample size. Comparing the variational method to Laplace's method, the variational method showed more negative bias when $\sigma^2 = 1$, and less bias when $J = 25$ and $\sigma^2 = 3$. Furthermore, the estimates stemming from the variational method showed overall more dispersion than the estimates stemming from Laplace's method.

With respect to the item parameters (Figures 4 and 5), both methods seemed to perform quite well when $J = 25$, except for the condition with $N = 500$ and $\sigma^2 = 1$, where the variational method showed a small shrinkage of the estimates towards zero, as can be inferred from the pattern in the boxplots: for $\beta_j = 2$ (leftmost boxplot), there was a negative bias, turning into a positive bias for $\beta_j = -2$ (rightmost boxplot). For $J = 5$, no consistent pattern could be discerned for Laplace's method, whereas the variational method showed a shrinkage towards zero for all conditions with $J = 5$. For the extreme item parameters ($\beta = 2$ and $\beta = -2$), the bias was about 10 to 20%. The dispersion was comparable for both methods under all conditions. Again, as could be expected, the dispersion was smaller for $N = 500$ than for $N = 100$. Furthermore, the dispersion was somewhat smaller when $\sigma^2 = 1$ compared to $\sigma^2 = 3$.

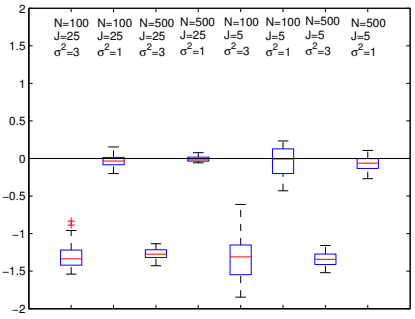
With respect to estimation time, we observed that the variational method was quite fast in all conditions whereas Laplace's method (as it is implemented in the NLMIXED procedure of SAS) slowed down considerably for the conditions with $J = 25$.

Fig. 3. Differences between estimated and true values of σ^2 for all eight experiments

Variational approximation



Laplace



5 Concluding remarks

We presented a variational estimation method for the mixed logistic regression model. The method is based on a lower bound approximation of the logistic function (Jaakkola and Jordan, 2000) and results in a considerable simplification of the maximization problem in that the approximated log-likelihood function no longer contains intractable integrals over the random effects. Jaakkola and Jordan (2000) used the lower bound approximation of the logistic function to approximate the posterior distribution of a logistic regression model formulated in a Bayesian framework. The variational estimation method we presented is a generalization of their approach in the sense that the mixed logistic regression model contains both fixed and random effects, whereas in a Bayesian framework, all parameters are considered to be random variables.

The results of the simulation study show that the variational method suffers from a negative bias in estimating the variance parameter(s) of the random effects. However, the bias decreases with an increasing number of outcomes obtained from the same cluster and its performance becomes comparable to Laplace's method. With respect to the fixed effects, it was also found that the performance of the variational method depends on the number of outcomes within the same cluster. For a small number of outcomes (i.e. 5), the variational method suffers from a shrinkage towards zero. This shrinkage was not or only to a minor degree observed for a larger number of outcomes (i.e. 25).

Together with the observation that estimation time increases more rapidly with an increasing number of outcomes for Laplace's method than for the variational method, the variational method seems to be a viable choice for estimating the fixed effects of the mixed logistic regression model when the number of outcomes within each cluster is sufficiently high.

Acknowledgements

Frank Rijmen was partly supported by the Fund for Scientific Research Flanders and Jiří Vomlel by the Ministry of Education of the Czech Republic under the projects nr. 1M0572 and nr. 2C06019.

Fig. 4. Differences between estimated and true values of $\boldsymbol{\beta}$ in experiments 5, 6, 7, and 8.

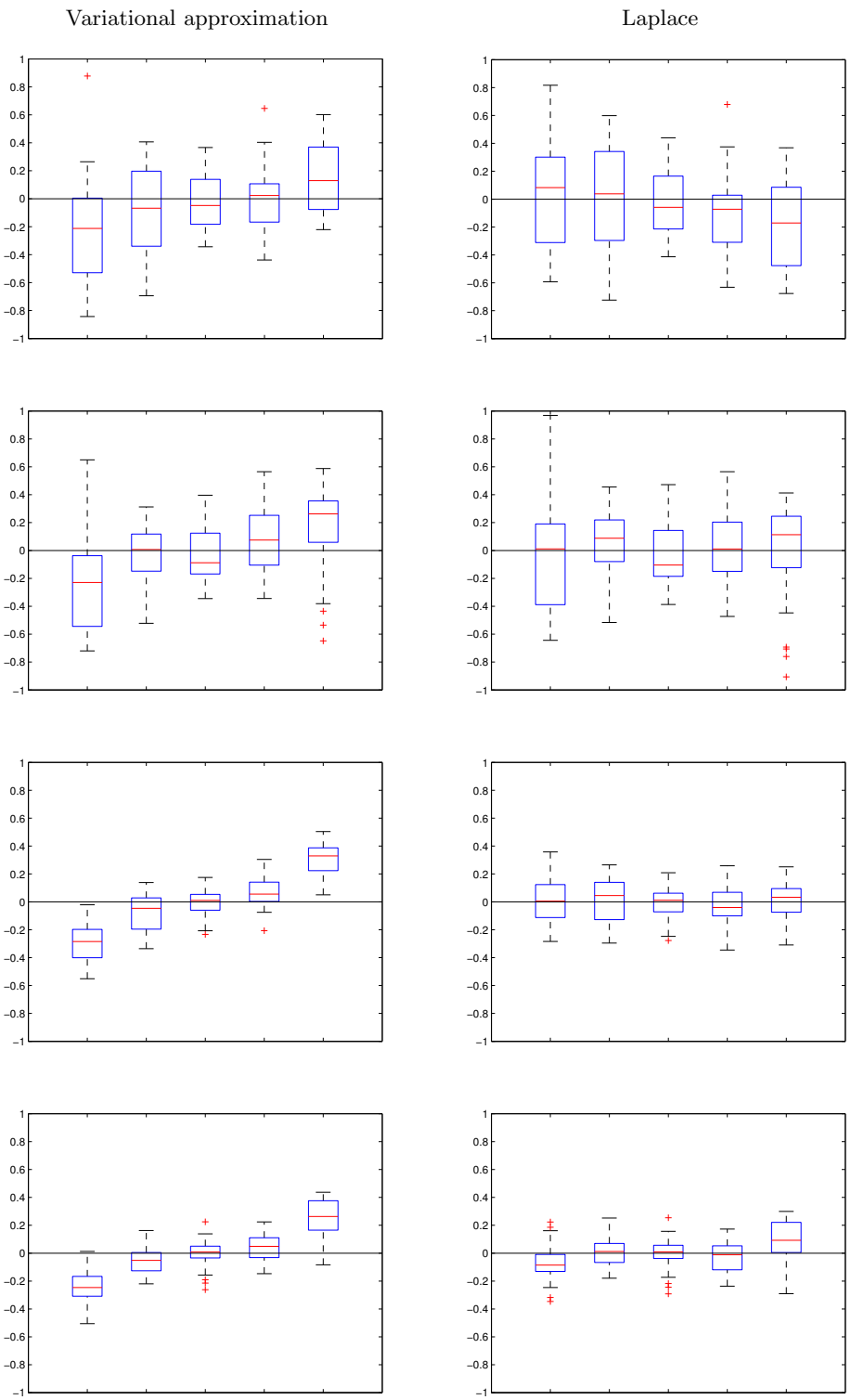
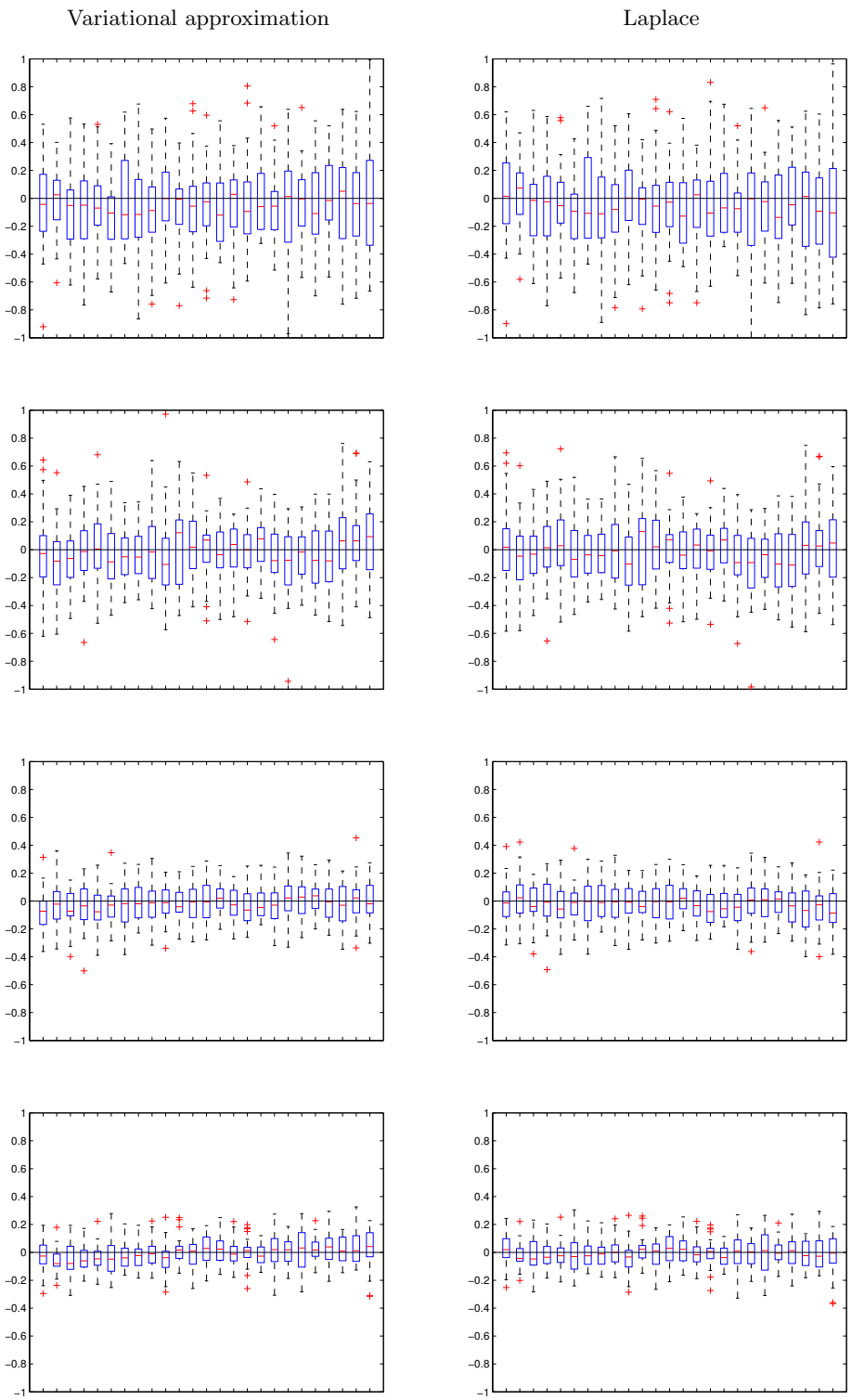


Fig. 5. Differences between estimated and true values of $\boldsymbol{\beta}$ in experiments 1, 2, 3, and 4.



Bibliography

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society 88*, 9–25.

Davidian, M. and D. M. Giltinan (1995). *Nonlinear Models for Repeated Measurement Data*. New York: Chapman & Hall.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B 39*, 1–38.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of longitudinal data* (Second ed.). Oxford University Press.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika 78*, 45–51.

Goldstein, H. and J. Rasbash (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A 159*, 505–513.

Hall, P., K. Humphreys, and D. M. Titterington (2002). On the adequacy of variational lin-ear bound functions for likelihood-based inference in Markovian models with missing values. *Journal of the Royal Statistical Society, B 64*, 549–564.

Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics & Computing 10*, 25–37.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. *Machine Learning 37*(2), 183–233.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Technical report, Danish Institute for Educational Research, Copenhagen.

Raudenbush, S. W., M. L. Yang, and M. Yosef (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate laplace approximation. *Journal of Computational and Graphical Statistics 9*, 141–157.

Rodriguez, G. and N. Goldman (1995). An assessment of estimation procedures for multilevel models with binary response. *Journal of the Royal Statistical Society, Series A 158*, 73–89.

SAS Institute Inc. (2005). *SAS OnlineDoc*. Cary, NC: SAS Institute Inc. Version 9.1, `http://support.sas.com/91doc/docMainpage.jsp`.

Saul, L. K., T. Jaakkola, and M. I. Jordan (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research 4*, 61–76.

Tierny, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*, 82–86.

Tipping, M. E. (1998). Probabilistic visualisation of high-dimensional binary data. In *Advances in Neural Information Processing Systems*, pp. 592–598. MIT Press.

Tuerlinckx, F., F. Rijmen, G. Verbeke, and P. De Boeck (2006). Statistical inference in generalized linear mixed models: A review. *British Journal for Mathematical and Statistical Psychology*. In press.

Verbeke, G. and G. Molenberghs (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.