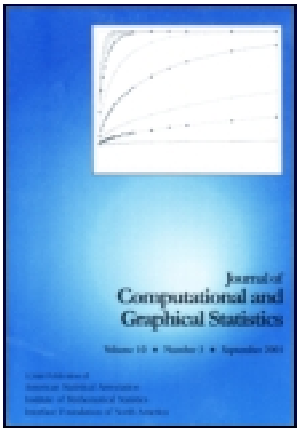

Journal of Computational and Graphical Statistics
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/ucgs20

GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ_1-Penalization

Jürg Schelldorfer , Lukas Meier , Peter Bühlmann [b] , AXA Winterthur [a] & ETH Zürich [c]

[a] CH-8400, Winterthur, Switzerland

[b] Seminar für Statistik

[c] CH-8092, Zürich, Switzerland

Accepted author version posted online: 08 Feb 2013.

GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using ℓ_1-Penalization

Jürg Schelldorfer, Lukas Meier and Peter Bühlmann *
AXA Winterthur and ETH Zürich

November 20, 2012

Abstract

We propose an ℓ_1-penalized algorithm for fitting high-dimensional generalized linear mixed models. Generalized linear mixed models (GLMMs) can be viewed as an extension of generalized linear models for clustered observations. Our Lasso-type approach for GLMMs should be mainly used as variable screening method to reduce the number of variables below the sample size. We then suggest a refitting by maximum likelihood based on the selected variables only. This is an effective correction to overcome problems stemming from the variable screening procedure which are more severe with GLMMs than for generalized linear models. We illustrate the performance of our algorithm on simulated as well as on real data examples. Supplemental materials are available online and the algorithm is implemented in the `R` package `glmmixedlasso`.

Key Words: coordinate gradient descent; Laplace approximation; random-effects model; variable selection.

1 Introduction

In recent years, high-dimensional linear regression models have been extensively studied. The most popular method to achieve sparse estimates is the Lasso (Tibshirani, 1996), which uses an

*Jürg Schelldorfer is Statistician (E-mail: juerg.schelldorfer@axa-winterthur.ch), AXA Winterthur, CH-8400 Winterthur, Switzerland, Lukas Meier is Senior Scientist (E-mail: meier@stat.math.ethz.ch), Peter Bühlmann is Professor (E-mail: buhlmann@stat.math.ethz.ch), Seminar für Statistik, ETH Zürich, CH-8092 Zürich, Switzerland.

ACCEPTED MANUSCRIPT

ℓ_1-penalty. The Lasso is not only attractive in terms of its statistical properties but also due to its fast computation solving a convex optimization problem. However, relatively few articles examine high-dimensional regression problems involving a non-convex loss function, i.e. Khalili and Chen (2007) and Städler et al. (2010) for Gaussian mixture models, Pan and Shen (2007) and Witten and Tibshirani (2010) for clustering and Witten and Tibshirani (2011) for linear discriminant analysis.

Generalized linear mixed models (McCullagh and Nelder, 1989; Breslow and Clayton, 1993; McCulloch and Searle, 2001; Molenberghs and Verbeke, 2005) are an extension of generalized linear models by adding random effects to the linear predictor in order to accommodate for clustered or overdispersed data. These models have received much attention in many applications such as biology, ecology, medicine, pharmaceutical science and econometrics. Available software packages (`lme4` in `R`, NLMIXED in SAS, among others) allow to fit a wide range of generalized linear mixed models.

In this paper we develop a method for high-dimensional generalized linear mixed models. It is based on a Lasso-type regularization with a cyclic coordinate descent optimization. Due to shrinkage introduced by ℓ_1-penalization, our approach performs in a first step variable screening, thereby selecting a set of candidate active variables. In other words, the proposed method primarily aims at reducing the dimensionality of the high-dimensional GLMM. In a second step, we perform refitting by maximum likelihood estimation to get accurate parameter estimates. The idea of such a two-stage approach has been used in linear models (Efron et al., 2004) and it is related to the adaptive Lasso (Zou, 2006) and the thresholded Lasso (Zhou, 2010; van de Geer et al., 2011). In fact, a two-stage approach is much more important than for linear models since shrinkage in GLMMs can have a severe effect on the estimation of variance components, see Sections 4 and 5.

To the best of our knowledge, there does not exist any literature devoted to truly high-dimensional generalized linear mixed models. Some papers focus on penalized variable selection procedures in generalized mixed models with low-dimensional data: we refer to Yang (2007), Ibrahim et al. (2010), Ni et al. (2010). Groll and Tutz (2012) have independently studied the same statistical problem and have also used a Lasso-type approach but with a focus on rather low-dimensional problems. Few papers focus on variable selection in generalized additive mixed models, for example Xue et al. (2010) and Lai et al. (2012). Schelldorfer et al. (2011) present statistical theory

ACCEPTED MANUSCRIPT

and an algorithm for high-dimensional Gaussian linear mixed models, where computation is much easier than in the generalized case.

The main contribution of the present paper is the construction and implementation of an efficient algorithm for ℓ_1-penalization in truly high-dimensional generalized linear mixed models, called the GLMMLasso. We use the Laplace approximation (Bates, 2011b) and combine it with efficient coordinate gradient descent methods (Tseng and Yun, 2009). Our algorithm is feasible for problems where the number of variables is in the thousands and taking advantage of sparsity with respect to dimensionality (i.e. only few active variables) is exploited by an active set strategy.

The rest of the article is organised as follows. In Section 2, we review the generalized linear mixed model and introduce the GLMMLasso estimator. In Section 3, we describe the details of the computational algorithm before advocating the two-stage GLMMLasso estimators in Section 4. In Section 5 and 6 we consider the performance of our methods on simulated and real data sets. The article concludes with a discussion in Section 7. Supplemental materials including additional simulation examples are available online.

2 Generalized linear mixed models and ℓ_1-penalized estimation

In this section, we first look at the classical GLMM setting where the number of observations is larger than the number of covariates, i.e. $p < n$. We closely follow Bates (2011a). Secondly, we consider the high-dimensional framework, i.e. $n \ll p$, and present the ℓ_1-penalized maximum likelihood estimator.

2.1 Model formulation

Suppose that the observations are not independent but grouped instead. Let $r = 1, \ldots, N$ be the grouping index and $j = 1, \ldots, n_r$ the jth outcome within group r. Denote by n the total number of observations, i.e. $n = \sum_{r=1}^{N} n_r$. Let $\boldsymbol{X}$ be the $n \times p$ fixed-effects design matrix, $\boldsymbol{Z}$ the $n \times q$ random-effects design matrix, $\boldsymbol{\mathcal{Y}}$ the n-dimensional random response vector and $\boldsymbol{\mathcal{B}}$ be the q-dimensional vector of random effects. We observe $\boldsymbol{y}$ of $\boldsymbol{\mathcal{Y}}$ whereas $\boldsymbol{\mathcal{B}}$ is unobserved. The generalized linear mixed model is specified by the unconditional distribution of $\boldsymbol{\mathcal{B}}$ and the conditional distribution of

$\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}$:

i) $\mathcal{Y}_i|\boldsymbol{\mathcal{B}} = \boldsymbol{b}$ are independent for $i = 1, \ldots, n$.

ii) The distribution of $\mathcal{Y}_i|\boldsymbol{\mathcal{B}} = \boldsymbol{b}$ belongs to the exponential family with density

$$\exp\left\{\phi^{-1}\Big(y_i\xi_i - b(\xi_i)\Big) + c(y_i, \phi)\right\},$$

where $b(.)$ and $c(., .)$ are known functions. ϕ is the dispersion parameter (known or unknown) and ξ_i is associated with the conditional mean $\mu_i := E[\mathcal{Y}_i|\boldsymbol{\mathcal{B}} = \boldsymbol{b}]$, i.e. $\xi_i = \xi_i(\mu_i)$.

iii) The conditional mean vector $\boldsymbol{\mu}$ depends on $\boldsymbol{b}$ through the known link function g and the linear predictor $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Zb}$, with $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\mu})$ componentwise. Here, $\boldsymbol{\beta}$ is the unknown p-dimensional parameter vector, called fixed effects, and $\boldsymbol{b}$ the unknown q-dimensional vector of random effects.

iv) $\boldsymbol{\mathcal{B}} \sim \mathcal{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$ where the covariance matrix $\boldsymbol{\Sigma}_\theta$ is parameterized by the unknown parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$. We assume that $\boldsymbol{\Sigma}_\theta$ is positive semidefinite, i.e. $\boldsymbol{\Sigma}_\theta \geq 0$. The dimensionality d is typically small, say $d \leq 10$.

By using $\boldsymbol{\mathcal{B}}$ and $\boldsymbol{\Sigma}_\theta$ in the definition above, we have already defined the random-effects structure of the GLMM. To be more precise, we have specified which variables have an additional random effect and how the structure of $\boldsymbol{\Sigma}_\theta$ looks like (e.g. multiple of the identity or diagonal). A discussion of how to find these structures is beyond the scope of this paper.

Let us write $\boldsymbol{\Sigma}_\theta$ in terms of its Cholesky decomposition $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Lambda}_\theta\boldsymbol{\Lambda}_\theta^T$ and introduce the (unobserved) random variable $\boldsymbol{\mathcal{U}}$ defined by $\boldsymbol{\mathcal{B}} := \boldsymbol{\Lambda}_\theta\boldsymbol{\mathcal{U}}$ where $\boldsymbol{\mathcal{U}} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{1}_q)$. Then the linear predictor $\boldsymbol{\eta}$ can be written as $\boldsymbol{\eta} = \boldsymbol{X\beta} + \boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}$. We estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and ϕ (if unknown) by the maximum likelihood method and predict the random effects $\boldsymbol{u}$.

2.2 Likelihood function

Employing the notation $\xi_i(\mu_i) = \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta})$, the likelihood function of a GLMM is given by the following expression:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) &= \int_{\mathbb{R}^q} \prod_{i=1}^{n} \left[\exp\left\{ \phi^{-1}\Big(y_i \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta})) \Big) + c(y_i, \phi) \right\} \right] \frac{1}{(2\pi)^{q/2}} \exp\Big\{ -\frac{1}{2} \|\boldsymbol{u}\|_2^2 \Big\} d\boldsymbol{u} \\ &= \frac{1}{(2\pi)^{q/2}} \int_{\mathbb{R}^q} \exp\left\{ \sum_{i=1}^{n} \Big(\frac{y_i \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\phi} + c(y_i, \phi) \Big) - \frac{1}{2} \|\boldsymbol{u}\|_2^2 \right\} d\boldsymbol{u}. \end{aligned} \tag{1}$$

In general, the integral (1) can not be worked out analytically and numerical approximations are required, see Skrondal and Rabe-Hesketh (2004), Molenberghs and Verbeke (2005) and Jiang (2007).

2.3 The GLMMLasso estimator

We now turn to the high-dimensional setting where the number of fixed-effect variables p is much larger than the number of observations n, i.e. we study the so-called $n \ll p$ setup.

Let us assume that the true underlying fixed-effects vector $\boldsymbol{\beta}_0$ is sparse in the sense that many coefficients of $\boldsymbol{\beta}_0$ are zero. To enforce sparsity of our estimator, we advocate a Lasso-type approach. This means that we add an ℓ_1-penalty for the fixed-effects vector $\boldsymbol{\beta}$ to the likelihood function. Thus, we are going to consider the following objective function:

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) = -2 \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) + \lambda \|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\lambda \geq 0$ is a regularization parameter. Appropriate choices for λ are discussed in Section 4.

We aim at estimating the fixed-effect parameter $\boldsymbol{\beta}$, the covariance parameter $\boldsymbol{\theta}$, and if unknown the dispersion parameter ϕ, by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi}) := \underset{\beta, \theta, \phi}{\arg\min}\, Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi). \tag{3}$$

We call (3) the GLMMLasso estimator. Since the likelihood function (1) comprises analytically intractable integrals (except for the Gaussian case), some approximations have to be used. We are going to illustrate the algorithm using the Laplace approximation. For GLMMs, it is accurate with low computational burden, as advocated by Bates (2011b). A thorough discussion of the accuracy

ACCEPTED MANUSCRIPT

and limitations of the Laplace approximation can be found in Joe (2008). Generally, the Laplace approximation is used to calculate integrals of the form

$$I = \int_{\mathbb{R}^q} e^{-S(\boldsymbol{u})} d\boldsymbol{u}, \tag{4}$$

where $S(\boldsymbol{u})$ is a known function of a q-dimensional variable $\boldsymbol{u}$. Let

$$\tilde{\boldsymbol{u}} = \arg\max_{u} -S(\boldsymbol{u}) \tag{5}$$

(i.e. $S'(\tilde{\boldsymbol{u}}) = 0$), then the Laplace approximation of I is given by

$$I \approx I^{LA} = (2\pi)^{q/2} |S''(\tilde{\boldsymbol{u}})|^{-1/2} e^{-S(\tilde{u})}. \tag{6}$$

The mode $\tilde{\boldsymbol{u}}$ in (5) is calculated by the penalized iterative least squares (PIRLS) algorithm. It is presented in Bates (2011b) and described in the supplemental materials. The PIRLS algorithm is related to the iterative reweighted least squares (IRLS) algorithm for obtaining the maximum likelihood estimator in generalized linear models.

It should be noted that $\tilde{\boldsymbol{u}}$ depends on $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and ϕ. From (1) and (6) we deduce that the Laplace approximation of the objective function $Q_\lambda(.)$ in (2) is

$$\begin{aligned} Q_\lambda^{LA}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) = & -2 \sum_{i=1}^{n} \left\{ \frac{y_i \xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta}, \boldsymbol{\theta}))}{\phi} + c(y_i, \phi) \right\} + \log |(\boldsymbol{Z}\boldsymbol{\Lambda}_\theta)^T \boldsymbol{W}_{\beta,\theta,\phi} (\boldsymbol{Z}\boldsymbol{\Lambda}_\theta) + \mathbf{1}_q| \\ & + \|\tilde{\boldsymbol{u}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \end{aligned} \tag{7}$$

where $\boldsymbol{W}_{\beta,\theta,\phi} = \operatorname{diag}^{-1}\left(\phi v(\mu_i(\boldsymbol{\beta}, \boldsymbol{\theta})) g'(\mu_i(\boldsymbol{\beta}, \boldsymbol{\theta}))^2\right)_{i=1}^n$ and $v(.)$ is the known conditional variance function (McCullagh and Nelder, 1989). The estimator (3) is then approximated by

$$(\hat{\boldsymbol{\beta}}^{LA}, \hat{\boldsymbol{\theta}}^{LA}, \hat{\phi}^{LA}) := \arg\min_{\beta,\theta,\phi} Q_\lambda^{LA}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi). \tag{8}$$

We call (8) the GLMMLasso^{LA} estimator. It is the approximation (8) to the objective function (3) that is optimized to obtain the parameter estimates. Moreover, we would like to emphasize that (8) is a non-convex function with respect to $(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi)$ consisting of a non-convex loss function and a convex penalty.

3 Computational algorithm

In this section, we present the computational algorithm to obtain the GLMMLasso^{LA} estimator (8). The algorithm is based on ideas from Tseng and Yun (2009) of the (block) coordinate gradient descent (CGD) method. The notion of the CGD algorithm is that we cycle through components of the full parameter vector $\boldsymbol{\psi} := (\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) \in \mathbb{R}^{p+d+1}$ and minimize the objective function $Q_\lambda^{LA}(.)$ only with respect to one parameter while keeping the other parameters fixed. In doing so we calculate a quadratic approximation and perform an indirect line search to ensure that the objective function decreases. (Block) CGD algorithms are used in Meier et al. (2008), Wu and Lange (2008), Friedman et al. (2010) and Breheny and Huang (2011) and are now extremely popular in high-dimensional penalized regression problems.

We first give an overview of the algorithm which solves minimization problem (8) exactly before considering an approximate algorithm which finds a solution close to the exact minimizer of (8). Finally, we present some details of the algorithm.

3.1 The exact GLMMLasso algorithm

We describe here an exact algorithm, called exact GLMMLasso (we notationally omit the involved Laplace approximation), for the Laplace approximated objective function in (8). Let us write (7) with a different notation to ease the presentation. For $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \phi) \in \mathbb{R}^{p+d+1}$, define the function

$$f(\psi) := -2\sum_{i=1}^{n}\left\{\frac{y_i\xi_i(\boldsymbol{\beta},\boldsymbol{\theta}) - b(\xi_i(\boldsymbol{\beta},\boldsymbol{\theta}))}{\phi} + c(y_i,\phi)\right\} + \log|(\boldsymbol{Z\Lambda}_\theta)^T\boldsymbol{W}_\psi(\boldsymbol{Z\Lambda}_\theta) + \mathbf{1}_q| + \|\tilde{\boldsymbol{u}}(\boldsymbol{\psi})\|_2^2.$$

Now (8) can be written as $\hat{\psi}_\lambda^{LA} = \arg\min_\psi Q_\lambda^{LA}(\psi) := f(\psi) + \lambda\|\boldsymbol{\beta}\|_1$. Let $\boldsymbol{e}_j$ be the jth unit vector and denote by (s) the sth iteration step. Moreover, we let

$$\boldsymbol{\beta}^{(s)} := (\beta_1^{(s)}, \ldots, \beta_p^{(s)})^T, \quad \boldsymbol{\theta}^{(s)} := (\theta_1^{(s)}, \ldots, \theta_d^{(s)})^T, \quad \phi^{(s)}$$

be the estimates of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and ϕ in the sth iteration. Using the notation

$$\begin{aligned}
\boldsymbol{\beta}^{(s,s-1,\beta_k)} &:= \left(\beta_1^{(s)},\ldots,\beta_{k-1}^{(s)},\beta_k,\beta_{k+1}^{(s-1)},\ldots,\beta_p^{(s-1)}\right)^T,\\
\boldsymbol{\theta}^{(s,s-1,\theta_l)} &:= \left(\theta_1^{(s)},\ldots,\theta_{l-1}^{(s)},\theta_l,\theta_{l+1}^{(s-1)},\ldots,\theta_d^{(s-1)}\right)^T,\\
\boldsymbol{\beta}^{(s,s-1;k)} &:= \left(\beta_1^{(s)},\ldots,\beta_{k-1}^{(s)},\beta_k^{(s-1)},\beta_{k+1}^{(s-1)},\ldots,\beta_p^{(s-1)}\right)^T,
\end{aligned}$$

the exact GLMMLasso algorithm is summarized in Algorithm 1.

Particularly in the high-dimensional setting, the calculation of the quadratic approximation requires a large amount of computing time. Therefore it is interesting to examine a much faster approximate algorithm.

3.2 The (approximate) GLMMLasso algorithm

In the exact Algorithm 1 above, we consider in step (1) b) the mode $\tilde{\boldsymbol{u}}$ as a function of the parameters, i.e. $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\boldsymbol{\beta},\boldsymbol{\theta},\phi)$. However, the calculation of the derivatives of $f(.)$ with respect to β_k is computationally intensive. This becomes a major issue in the high-dimensional setting where a substantial amount of computing time is allocated to this particular part of the algorithm. In addition, the exact GLMMLasso algorithm requires a large number of outer iterations s. To attenuate these difficulties, we propose a slightly modified version of Algorithm 1. We suggest performing the quadratic approximation and the inexact line search while considering $\tilde{\boldsymbol{u}}$ as fixed and not depending on β_k. Denoting by $f(.|\tilde{\boldsymbol{u}})$ the function $f(.)$ for which $\tilde{\boldsymbol{u}}$ is considered as fixed, the (approximate) GLMMLasso algorithm is given in Algorithm 2:

We illustrate in the supplemental materials that the approximate GLMMLasso algorithm speeds up remarkably without loosing that much accuracy. Additionally, the approximation emphasizes the importance of a refitting as advocated in the next section.

3.3 Convergence behaviour and details of the GLMMLasso algorithm

Numerical convergence. The convergence of the exact GLMMLasso algorithm to a stationary point can be proofed using the results presented in Tseng and Yun (2009). It is worth pointing out that in the low-dimensional framework, the exact GLMMLasso algorithm with $\lambda = 0$ (no penal-

Algorithm 1 *Exact GLMMLasso algorithm*

(0) Choose a starting value $\psi^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}, \phi^{(0)})$.

Repeat for $s = 1, 2, \ldots$

(1) *(fixed-effect parameter optimization)*
For $k = 1, \ldots, p$

a) *(Laplace approximation)*
Calculate the Laplace approximation

$$Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big).$$

b) *(Quadratic approximation and inexact line search)*

i) Approximate the second derivative

$$\frac{\partial^2}{\partial \beta_k^2} f\Big(\boldsymbol{\beta}^{(s,s-1,\beta_k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big)\Big|_{\beta_k=\beta_k^{(s-1)}}$$

by $h_k^{(s)} > 0$ as described in the subsection below.

ii) Calculate the descent direction $d_k^{(s)} \in \mathbb{R}$

$$d_k^{(s)} := \arg\min_d \Big\{ f\Big(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big) + \frac{\partial}{\partial \beta_k} f\Big(\boldsymbol{\beta}^{(s,s-1,\beta_k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big)\Big|_{\beta_k=\beta_k^{(s-1)}} d + \frac{1}{2} d^2 h_k^{(s)} + \lambda \|\boldsymbol{\beta}^{(s,s-1;k)} + d\boldsymbol{e}_k\|_1 \Big\}.$$

iii) Choose a step size $\alpha_k^{(s)} > 0$ and set $\boldsymbol{\beta}^{(s,s-1;k+1)} = \boldsymbol{\beta}^{(s,s-1;k)} + \alpha_k^{(s)} d_k^{(s)} \boldsymbol{e}_k$ such that

$$Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k+1)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big) \le Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big).$$

(2) *(Covariance parameter optimization)*
For $l = 1, \ldots, d$

$$\theta_l^{(s)} = \arg\min_{\theta_l} Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s)}, \boldsymbol{\theta}^{(s,s-1;\theta_l)}, \phi^{(s-1)}\Big).$$

(3) *(Dispersion parameter optimization)*

$$\phi^{(s)} = \arg\min_{\phi} Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s)}, \boldsymbol{\theta}^{(s)}, \phi\Big).$$

until convergence.

ization) gives the same results as the function `glmer` in the `R` package `lme4`.

(0) Starting value $\psi^{(0)}$. As starting value for $\boldsymbol{\beta}$, we fit a generalized linear model with the Lasso where the regularization parameter is chosen by cross-validation. The initial values for $\boldsymbol{\theta}$ and ϕ are then calculated using steps (2) and (3) in Algorithm 1 and 2.

Algorithm 2 *(Approximate) GLMMLasso algorithm*

Denote by $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)})$. Replace in Algorithm 1 i) - iii) by

i') Approximate the second derivative

$$\frac{\partial^2}{\partial \beta_k^2} f\Big(\boldsymbol{\beta}^{(s,s-1,\beta_k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big|\tilde{\boldsymbol{u}}\Big)\Big|_{\beta_k=\beta_k^{(s-1)}}$$

by $h_k^{(s)} > 0$ as described in the subsection below.

ii') Calculate the descent direction $d_k^{(s)} \in \mathbb{R}$

$$d_k^{(s)} := \arg\min_d \Big\{ f\Big(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big|\tilde{\boldsymbol{u}}\Big) + \frac{\partial}{\partial \beta_k} f\Big(\boldsymbol{\beta}^{(s,s-1,\beta_k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big|\tilde{\boldsymbol{u}}\Big)\Big|_{\beta_k=\beta_k^{(s-1)}} d$$
$$+ \frac{1}{2} d^2 h_k^{(s)} + \lambda \|\boldsymbol{\beta}^{(s,s-1;k)} + d\boldsymbol{e}_k\|_1 \Big\}.$$

iii') Choose a step size $\alpha_k^{(s)} > 0$ and set $\boldsymbol{\beta}^{(s,s-1;k+1)} = \boldsymbol{\beta}^{(s,s-1;k)} + \alpha_k^{(s)} d_k^{(s)} \boldsymbol{e}_k$ such that

$$Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k+1)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big|\tilde{\boldsymbol{u}}\Big) \le Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big|\tilde{\boldsymbol{u}}\Big).$$

i) Choice of $h_k^{(s)}$. For $h_k^{(s)}$ we choose the kth diagonal element of the Fisher information of a generalized linear model. Hence we use the second derivative of the first summand in (7). We set $c_{min} \le h_k^{(s)} \le c_{max}$ for positive constants c_{min} and c_{max} (e.g. $c_{min} = 10^{-5}$ and $c_{max} = 10^5$) in order that the algorithm converges (Tseng and Yun, 2009).

ii) Calculation of $d_k^{(s)}$. The value $d_k^{(s)}$ is the minimizer of the quadratic approximation of the objective function $Q_\lambda^{LA}(.)$ and analytically given by (Tseng and Yun, 2009)

$$d_k^{(s)} = \begin{cases} \text{median}\left(\dfrac{\lambda - \partial/\partial\beta_k f_{\beta_k}}{h_k^{(s)}}, -\beta_k, \dfrac{-\lambda - \partial/\partial\beta_k f_{\beta_k}}{h_k^{(s)}}\right) & \text{if } \beta_k \text{ penalized} \\ -\frac{\partial/\partial_{\beta_k} f_{\beta_k}}{h_k^{(s)}} & \text{otherwise,} \end{cases} \tag{9}$$

where $f_{\beta_k} = f(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)})$ in Algorithm 1 and $f_{\beta_k} = f(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\big|\tilde{\boldsymbol{u}})$ in Algorithm 2.

iii) Choice of $\alpha_k^{(s)}$. The step length $\alpha_k^{(s)}$ is chosen such that the objective function $Q_\lambda^{LA}(.)$ decreases. We suggest to use the Armijo rule, which is defined for Algorithm 1 as follows (and correspondingly for Algorithm 2 with fixed $\tilde{\boldsymbol{u}}$):

Armijo rule: Choose $\alpha_k^{init} > 0$ *and let* $\alpha_k^{(s)}$ *be the largest element of* $\{\alpha_k^{init}\delta^l\}_{l=0,1,2,..}$ *satisfying*

$$Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k)} + \alpha_k^{(s)} d_k^{(s)} \boldsymbol{e}_k, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big) \leq Q_\lambda^{LA}\Big(\boldsymbol{\beta}^{(s,s-1;k)}, \boldsymbol{\theta}^{(s-1)}, \phi^{(s-1)}\Big) + \alpha_k^{(s)} \varrho \triangle^k$$

where $\triangle^k := \partial/\partial\beta_k f_{\beta_k} d_k^{(s)} + \gamma(d_k^{(s)})^2 h_k^{(s)} + \lambda\|\boldsymbol{\beta}^{(s,s-1;k)} + d_k^{(s)}\boldsymbol{e}_k\|_1 - \lambda\|\boldsymbol{\beta}^{(s,s-1;k)}\|_1$.

The choice of the constants comply with the suggestions in Bertsekas (1999), e.g. $\alpha_k^{init} = 1$, $\delta = 0.5$, $\varrho = 0.1$ and $\gamma = 0$.

Active Set Algorithm. If we assume that the true fixed-effect parameter $\boldsymbol{\beta}_0$ is sparse in the sense that many elements are zero, we can reduce the computing time remarkably by using an active set algorithm. This is also used in Meier et al. (2008) and Friedman et al. (2010). In particular, we only cycle through all p coordinates every Dth iteration, otherwise only through the current active set $S(\hat{\boldsymbol{\beta}}^{(s-1)}) = \{k : \hat{\beta}_k^{(s-1)} \neq 0\}$. Typical values for D are 5 and 10.

An implementation of the algorithm is given in the `R` package `glmmixedlasso` and will be made available on R-Forge (`http://r-forge.R-project.org/`).

4 The two-stage GLMMLasso^{LA} estimator(s)

From the soft-thresholding property of the Lasso in linear models (Tibshirani, 1996) and in Gaussian linear mixed models (Schelldorfer et al., 2011), the fixed-effect estimate $\hat{\boldsymbol{\beta}}$ is biased towards zero. In some generalized linear mixed models the estimate of the covariance parameter $\boldsymbol{\theta}$ is biased, too. To mitigate these bias problems and the approximation error induced by using the approximate GLMMLasso algorithm, we advocate a two-stage procedure. The first step aims at estimating a candidate set of predictors $\hat{S}$ and can be seen as a variable screening procedure. The purpose of the second step is a more unbiased estimation of the parameters using unpenalized maximum likelihood (ML) estimation based on the selected variables $\hat{S}$ from the first step. The proposed two-stage GLMMLasso algorithm is summarized in Algorithm 3:

Algorithm 3 *Two-stage GLMMLasso algorithm*

Stage 1: Compute the GLMMLasso^{LA} estimate (8) and the set $\hat{S}$.

Stage 2: Perform unpenalized ML estimation.

In the next subsections, we are going to discuss the specification of the set of variables $\hat{S}$. We propose two methods from the high-dimensional linear regression framework, and we do not consider the adaptive Lasso (Zou, 2006).

4.1 The GLMMLasso^{LA}-MLE hybrid estimator

The LARS-OLS hybrid estimator was examined in Efron et al. (2004) and also used in Meinshausen and Bühlmann (2006) and Meier et al. (2008). In our context, it becomes a two-stage procedure where the model is refitted including only the covariates with a nonzero fixed-effect coefficient in $\hat{\boldsymbol{\beta}}_{init}$, where $(\hat{\boldsymbol{\beta}}_{init}, \hat{\boldsymbol{\theta}}_{init}, \hat{\phi}_{init})$ denotes the initial estimate from (8). More specifically, choose $\hat{S} = \hat{S}_{init} := \{k : |\hat{\beta}_{k,init} \neq 0\}$. Then the GLMMLasso^{LA}-MLE hybrid estimator is given by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi})_{hybrid} := \arg\min_{\beta_{\hat{S}_{init}}, \theta, \phi} -2 \log L(\boldsymbol{\beta}_{\hat{S}_{init}}, \boldsymbol{\theta}, \phi), \tag{10}$$

where for $S \subseteq \{1, \ldots, p\}$, $(\boldsymbol{\beta}_S)_k = \beta_k$ if $k \in S$ and $(\boldsymbol{\beta}_S)_k = 0$ if $k \notin S$.

4.2 The thresholded GLMMLasso^{LA} estimator

The thresholded Lasso with refitting in high-dimensional linear regression models was examined in van de Geer et al. (2011) and Zhou (2010). We define the set $\hat{S}_{thres}$ to be the set of variables which have initial fixed-effect coefficients larger than some threshold $\lambda_{thres} > 0$, i.e. we choose $\hat{S} = \hat{S}_{thres} := \{k : |\hat{\beta}_{k,init}| > \lambda_{thres}\}$. The thresholded GLMMLasso^{LA} estimator is then defined by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi})_{thres} := \arg\min_{\beta_{\hat{S}_{thres}}, \theta, \phi} -2 \log L(\boldsymbol{\beta}_{\hat{S}_{thres}}, \boldsymbol{\theta}, \phi). \tag{11}$$

The thresholded GLMMLasso^{LA} estimator involves another regularization parameter λ_{thres}, which is determined by minimizing an information criterion presented in the next subsection.

4.3 Selection of the regularization parameters

Estimators (8), (10) and (11) require the choice of the regularization parameters λ and λ_{thres}, respectively. We propose to use the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC), defined by

$$c_{n,\lambda} = -2 \log L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\phi}) + a(n) \cdot \hat{df}_{\lambda} \tag{12}$$

where $a(n) = \log(n)$ for the BIC and $a(n) = 2$ for the AIC. Here, $\hat{df}_{\lambda} = |\{1 \leq k \leq p : \hat{\beta}_k \neq 0\}| + \dim(\hat{\boldsymbol{\theta}})$ is the sum of of the number of nonzero fixed-effect coefficients and the number of covariance parameters. The first summand is motivated by the work of Zou et al. (2007). The second summand is the approach of Bates (2010), who proposes that in the classical generalized mixed effects model the degrees of freedom are given by the number of unconstrained optimization parameters. Based on our empirical experience, we suggest for the estimators (8) and (10) the BIC, whereas for (11) we advocate using the AIC (allowing for a larger number of variables) to select λ first and then, sequentially, the BIC to select λ_{thres}. We will compare the performance of the three estimators in the next sections.

5 Simulation Study

In this section we assess the performance of the GLMMLasso^{LA} estimators (8), (10) and (11). We compare them with appropriate Lasso, maximum likelihood (ML) and Penalized Quasi-Likelihood (PQL, Breslow and Clayton (1993)) methods.

In the main text, we only present simulation results for the high-dimensional logistic mixed model. Simulation studies for the low-dimensional logistic and the Poisson mixed model are included in the supplementary material. At the end of this section, we compare the GLMMLasso^{LA} estimates in a situation where the number of noise variables grows successively.

First of all, let us summarize some general conclusions drawn from real data analysis and the simulation studies:

a) The variable screening performance of the GLMMLasso algorithm is not only attractive for the high-dimensional setting, but also for low-dimensional data with a relatively large number of variables (say $p > 20$).

b) The GLMMLasso algorithm is numerically as stable as standard `R` functions like `glmer` (Bates, 2010) or `glmmPQL` (Breslow and Clayton, 1993; Venables and Ripley, 2002) when $p < n$. On the other hand, `glmpath` (Park and Hastie, 2007) and `glmnet` (Friedman et al., 2010) may fail to converge when high-dimensional models are misspecified.

c) The main difference between the logistic and the Poisson mixed model is the shrinkage of the covariance parameter estimates of the GLMMLasso^{LA} estimator. These estimates are severely biased in logistic mixed models, in contrast to the Poisson mixed model. Further differences between these two classes are summarized in the supplemental materials.

d) The number of iterations s substantially differs between the classes of generalized linear mixed models and the data set.

5.1 Preview for the logistic mixed model

In this section we confine the discussion to the logistic mixed model because it is viewed as the most challenging model within the class of generalized linear mixed models (Molenberghs and Verbeke, 2005; Jiang, 2007). As an overview, let us sum up the main findings from the simulation study in the logistic mixed model:

i) The GLMMLasso^{LA} estimate from (8) of the covariance parameter $\boldsymbol{\theta}$ is notably biased. In other words, adding an ℓ_1-penalty does not only shrink the fixed effects estimate $\hat{\boldsymbol{\beta}}$, but also the covariance parameter estimate $\hat{\boldsymbol{\theta}}$.

ii) In the high-dimensional settings, the GLMMLasso^{LA}-MLE hybrid estimator (10) performs better in terms of parameter estimation accuracy than the thresholded GLMMLasso^{LA} estimator (11).

iii) The more random effects, the more important it is to use the GLMMLasso^{LA} for variable screening (instead of a Lasso ignoring the grouping structure).

iv) The number of total iterations s needed is small, often about 15 iterations.

5.2 High-dimensional logistic mixed model

In all subsequent simulation schemes (including the supplemental materials), we restrict ourselves to the case where the number of observations per cluster is equal, i.e. $n_r = n_C$ for $r = 1, \ldots, N$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{V}$ with pairwise correlation $\boldsymbol{V}_{kk'} = \rho^{|k-k'|}$ and $\rho = 0.2$. Denote by $\boldsymbol{\beta}_0$ the true fixed effects (wherein $(\boldsymbol{\beta}_0)_1$ is the intercept) and by s_0 the true number of nonzero fixed-effect coefficients.

For the logistic mixed models, the intercept and the first covariate have independent random effects with different variance parameters. In particular, $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and covariance matrix $\boldsymbol{\Sigma}_\theta = \mathrm{diag}(\theta_1^2, , \ldots, \theta_1^2, \theta_2^2, \ldots, \theta_2^2) \in \mathbb{R}^{2N}$, i.e. $q = 2N$. We investigate the following two examples in the high-dimensional setting:

H_1: $N = 40$, $n_C = 10$, $n = 400$, $p = 500$, $\theta_1^2 = \theta_2^2 = 1$ and $s_0 = 5$ with $\boldsymbol{\beta}_0 = (0.1, 1, -1, 1, -1, 0, \ldots, 0)^T$.

H_2: $N = 50$, $n_C = 10$, $n = 500$, $p = 1500$, $\theta_1^2 = \theta_2^2 = 1$ and $s_0 = 5$ with $\boldsymbol{\beta}_0 = (0.1, 1, -1, 1, -1, 0, \ldots, 0)^T$.

The fitted models are all correctly specified. Hereafter, we denote by *oracle* the ML estimate of the model which includes only the variables from the true active set. Let *glmmlasso*, *hybrid glmmlasso* and *thres glmmlasso* be the GLMMLasso^{LA} estimates (8), (10) and (11), respectively. We compare the GLMMLasso^{LA} methods with the standard Lasso for generalized linear models (which ignore the grouping structure). For that purpose we use the *glmpath* algorithm (Park and Hastie, 2007) and the BIC as variable selection criterion. Then, let *hybrid glmpath* and *thres glmpath* be the two-stage procedures based on *glmpath* (without random effects).

The results in the form of median and rescaled median absolute deviation (in parentheses) over 100 simulation runs are shown in Table 1. There, $|S(\hat{\boldsymbol{\beta}})|$ denotes the cardinality of the estimated active set and TP is the number of true positives (selected variables which are in the true active set). SE is the squared error of the fixed-effect coefficients, i.e. SE$= ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2^2$.

Comparing the cardinality of the active set, we see that *thres glmmlasso* and *thres glmpath* have much larger active sets than *glmmlasso* and *glmpath*, respectively. This is largely due to the fact that we employ the AIC in the first and the BIC in the second stage. This is outweighed by the advantage that on average (not shown), the true effects are predominantly included in *thres glmmlasso*. The active set of *glmmlasso* is slightly smaller than that of *glmpath*. And yet, the number of TP is similar as for *glmpath*. Hence, we conclude that the existence of random effects does affect the variable selection performance of *glmpath*.

Concerning covariance parameter estimation, we read off from the table that $\hat{\theta}_1^2$ and $\hat{\theta}_2^2$ are seriously biased for *glmmlasso*. This motivates the usage of a two-stage procedure. The table suggests that the hybrid and the thresholded procedures have improved estimation accuracy of the random effects parameters compared to their original counterparts.

Looking at the fixed-effect parameter estimation accuracy, the simulation study reveals that the *glmmlasso* estimates are less biased than the corresponding *glmpath* estimates, resulting in lower squared error. And the same holds for *hybrid glmmlasso* and *hybrid glmpath*. The fixed-effect parameter estimates of *thres glmmlasso* and *thres glmpath* perform inadequately compared to their *hybrid* counterparts. As marked by an asterisk in the table, β_2 is not subject to penalization for the GLMMLasso^{LA} estimator since this variable has a random effect (Schelldorfer et al., 2011). Thus the bias of the estimate is much smaller than for the other fixed-effect coefficients.

To sum up the simulation study, we first conclude that *hybrid glmmlasso* outperforms *thres glmmlasso* in terms of parameter estimation accuracy, with similar performance regarding true positives. Second, *glmmlasso* procedures do outperform *glmpath* procedures as variable screening methods. Of course, *glmpath* is fitting a wrong model without random effects.

5.3 Logistic mixed model with a growing number of noise covariates

Here, we assess the performance of *glmmlasso* and *hybrid glmmlasso* when the number of noise variables grows successively. In the low-dimensional setting, we compare them with the ML estimate computed by the `R` function `glmer` (denoted by *glmer*). In addition, let *p-glmer* be the method which performs variable selection in the following way: Eliminate consecutively (backward selection) all variables with a p-value larger than 5% until the final model is attained comprising only significant variables. We compare these four methods in terms of their performance of twice the negative out-of-sample log-likelihood. Let us fix the following random intercept model design: $n = 400$, $N = 40$, $n_C = 10$, $\theta^2 = 1$, $\boldsymbol{\beta}_0 = (0, 1, -1, 1, -1)$. We start with $p = 5$ (no noise variables) and raise the number of variables to $p = 65$. The results over 50 simulation runs are depicted in Figure 1.

The figures show that the negative out-of-sample log-likelihood values for *glmer* grow polynomial whereas the likelihoods for the other methods remain fairly constant. The increase in *glmer* stems from the fact that it overfits the model for a growing number of covariates. When focusing on the figures in more detail, we read off that the negative log-likelihood of *glmmlasso* increases slightly for larger p whereas the negative log-likelihood of *hybrid glmmlasso* remains stable. The rationale for this small increase in *glmmlasso* is that the more noise covariates, the larger the optimal λ, and henceforth the larger the shrinkage of the fixed effects. And this leads to the increase of the out-of-sample log-likelihood. *hybrid glmmlasso* (and also *thres glmmlasso*) overcomes this problem and leads to a stable out-of-sample log-likelihood irrespective of p.

5.4 Correlated Random Effects

Both from a methodological and an implementational point of view it is conceptually possible to use correlated random effects. As an illustration we use the logistic mixed model H_1 with correlated random effects (with unstructured covariance matrix) where we use a correlation of $\rho = 0.5$ between the two random effects. The corresponding results are illustrated in Table 2. The results are very similar to the uncorrelated case. However, the bias of the correlation estimate seems to be less severe than the bias of the variance components.

6 Illustration

In this section we illustrate the proposed GLMMLasso^{LA} estimators for Poisson regression on an extended real data set with count data.

Data description. We consider the epilepsy data from Thall and Vail (1990) which were also analyzed by Breslow and Clayton (1993). The data were obtained from a randomized clinical trial of 59 patients with epilepsy, comparing a new drug (Trt=1) with placebo (Trt=0). The response variable consists of counts of epileptic seizures during the two weeks before each of four clinic visits (V4=1 for fourth visit, 0 otherwise). Further covariates in the analysis are the logarithm of age (Age), the logarithm of 1/4 the number of baseline seizures (Base) and the interaction of Base and Trt (Base x Trt). The main question of interest is whether taking the new drug reduces the number of epileptic seizures compared with placebo. In order to assess the performance of the proposed procedure with high-dimensional data, we add $U(-1, 1)$ distributed noise predictors to get a data set with $n = 236$, $N = 59$, $n_r = 4$ for $r = 1, \ldots, N$ and $p = 4000$. All predictors are standardized to have mean zero and standard deviation one.

Model. Model III in Breslow and Clayton (1993) is a two level GLMM (Bates, 2010), which is an extension of the single level GLMM introduced in Section 2 for more than one grouping variable. The model consists of two independent random intercept effects. One for subject (level 1, index r) and one for observation (level 2, index j). Let θ^2_{sub} and θ^2_{obs} be the corresponding variance parameters. Then the linear predictor can be written as

$$\log(\mu_{rj}) = \eta_{rj} = \boldsymbol{x}_{rj}^T\boldsymbol{\beta} + \theta_{sub}u_r + \theta_{obs}u_{rj} \quad r = 1, \ldots, 59, \quad j = 1, \ldots, 4.$$

Results. The results of the analysis are presented in Table 3. In the first column we show the estimates for Model III without performing variable selection. There, Intercept, Base and Trt are significant at the 5% level (indicated by †). If we perform backward selection using the BIC, we end up with a model including Intercept and Base only. And this model coincides with the one selected by *glmmlasso*. *Hybrid glmmlasso* overcomes the bias problems of *glmmlasso* and it yields a better model in terms of the BIC. *Thres glmmlasso* includes additional noise variables, thereby achieving the smallest BIC score for all models under consideration. Comparing *hybrid glmmlasso* and *thres glmmlasso*, the table suggests that the additional covariates in the latter model reduce the variability while keeping the fixed-effect estimates unaltered.

7 Concluding Remarks

We address the problem of estimating high-dimensional generalized linear mixed models (GLMMs). While low-dimensional generalized linear mixed models (Bates, 2010) and high-dimensional generalized linear models (van de Geer, 2008) have been extensively studied in recent years, little attention has been devoted to high-dimensional GLMMs. We provide an efficient algorithm for the

ACCEPTED MANUSCRIPT

ℓ_1-penalized maximum likelihood estimator, called GLMMLasso. It is based on the Laplace approximation, coordinatewise optimization and a speeding up approximation. The method should be typically used as a screening procedure to estimate a small set of important variables. We propose refitting by maximum likelihood to get accurate parameter estimates. The second stage is much more important than for linear models, because ℓ_1-shrinkage can lead to severe bias problems for the estimation of the variance components. Our work is primarily a contribution addressing the numerical challenges of performing high-dimensional variable selection and parameter estimation in nonlinear mixed-effects models involving a non-convex loss function. An implementation of the algorithm can be found in our `R` package `glmmixedlasso`. It will be made available on R-Forge.

SUPPLEMENTAL MATERIALS

All the following supplemental files can be obtained as a single zip file online (glmmlasso.zip):

Appendices: Details of the PIRLS algorithm, the comparison of the exact and approximate GLMMLasso algorithm and additional simulation studies. (glmmlasso_sm.pdf)

Data set: The extended epilepsy data set used in Section 6. (epilepsy.txt)

R-package for GLMMLasso : R-package `glmmixedlasso` containing code to perform the GLMMLasso algorithm. (glmmixedlasso-0.1-2.tar.gz)

ACKNOWLEDGEMENTS

The research is supported in part by the Swiss National Science Foundation (grant no. 20PA21-120043/1, “Forschergruppe FOR 916”). The authors thank the members of the DFG-SNF Forschergruppe 916 for many stimulating discussions. In particular we would like to thank Stephan Dlugosz from the ZEW Mannheim for insisting on studying this particular kind of problem. Moreover, the authors thank the associate editor and two referees for their helpful comments.

References

Bates, D. M. (2010). lme4: Mixed-effects Modeling with R. *Available at http://lme4.r-forge.r-project.org/book/*.

Bates, D. M. (2011a). Computational Methods for Mixed Models. *Available at http://cran.r-project.org/web/packages/lme4/vignettes/Theory.pdf*.

Bates, D. M. (2011b). Linear Mixed Model Implementation in lme4. *Available at http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf*.

Bertsekas, D. P. (1999). *Nonlinear Programming*. Belmont: Athena Scientific.

Breheny, P. and Huang, J. (2011). Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *The Annals of Applied Statistics*, 5:232–253.

Breslow, N. and Clayton, D. G. (1993). Approximate Inference in Generealized Linear Mixed Models. *Journal of the American Statistical Association*, 88:9–25.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32:407–499.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1).

Groll, A. and Tutz, G. (2012). Variable Selection for Generalized Linear Mixed Models by L_1-Penalized Estimation. *Statistics and Computing*. To appear.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2010). Fixed and Random Effects Selection in Mixed Effects Models. *Biometrics*, 67:495–503.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer.

ACCEPTED MANUSCRIPT

Joe, H. (2008). Accuracy of Laplace Approximation for Discrete Response Mixed Models. *Computational Statistics & Data Analysis*, 52:5066–5074.

Khalili, A. and Chen, J. (2007). Variable Selection in Finite Mixture of Regression Models. *Journal of the American Statistical Association*, 102:1025–1038.

Lai, R., Huang, H.-C., and Lee, T. (2012). Fixed and Random Effects Selection in Nonparametric Additive Mixed Models. *Electronic Journal of Statistics*, 6:810–842.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society, Ser. B*, 70:53–71.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional Graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:1436–1462.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

Ni, X., Zhang, D., and Zhang, H. H. (2010). Variable Selection for Semiparametric Mixed Models in Longitudinal Studies. *Biometrics*, 66:79–88.

Pan, W. and Shen, X. (2007). Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Research*, 8:1145–1164.

Park, M. and Hastie, T. (2007). L1-regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society, Ser. B*, 69:659–677.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization. *The Scandinavian Journal of Statistics*, 38:197–214.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling*. Boca Raton: Chapman & Hall/CRC.

Städler, N., Bühlmann, P., and van de Geer, S. (2010). l_1-Penalization for Mixture Regression Models (with discussion). *Test*, 19:209–285.

Thall, P. and Vail, S. (1990). Some Covariance Models for Longitudinal Count Data With Overdispersion. *Biometrics*, 46:657–671.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Ser. B*, 58:267–288.

Tseng, P. and Yun, S. (2009). A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization. *Mathematical Programming, Ser. B*, 117:387–423.

van de Geer, S. (2008). High-dimensional Generalized Linear Models and the Lasso. *The Annals of Statistics*, 36:614–645.

van de Geer, S., Bühlmann, P., and Zhou, S. (2011). The Adaptive and the Thresholded Lasso for Potentially Misspecified Models (and a Lower Bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Witten, D. M. and Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105:713–726.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, Ser. B*, in press.

Wu, T. and Lange, K. (2008). Coordinate Descent Algorithms for Lasso Penalized Regression. *The Annals of Applied Statistics*, 2:224–244.

Xue, L., Qu, A., and Zhou, J. (2010). Conistent Model Selection for Marginal Generalized Additive Model for Correlated Data. *Journal of the American Statistical Association*, 105:1517–1530.

ACCEPTED MANUSCRIPT

Yang, H. (2007). *Variable Selection Procedures for Generalized Linear Mixed Models in Longitudinal Data Analysis*. PhD thesis, North Carolina State University.

Zhou, S. (2010). Thresholded Lasso for High Dimensional Variable Selection and Statistical Estimation. *arXiv Preprint arXiv:1002.1583v2*.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the ”Degrees of Freedom” of the Lasso. *The Annals of Statistics*, 35:2173–2192.

ACCEPTED MANUSCRIPT

Table 1: *Simulation results (medians) for the logistic mixed models H_1 and H_2 (rescaled median absolute deviations in parentheses). A * means that the corresponding coefficient is not subject to penalization in the GLMMLasso^{LA} estimate.*

Model	Method	$\lvert S(\hat{\boldsymbol{\beta}})\rvert$	TP	$\hat{\theta}_1^2$	$\hat{\theta}_2^2$	$\hat{\beta}_1^*$	$\hat{\beta}_2^*$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	SE
True		5	5	1	1	0.1	1	-1	1	-1	
H_1	oracle	5	5	0.85	0.86	0.07	1.04	-0.99	0.98	-1.01	0.14
		(0)	(0)	(0.4)	(0.59)	(0.2)	(0.25)	(0.22)	(0.18)	(0.14)	(0.088)
	glmmlasso	6	5	0.38	0.37	0.06	0.66	-0.3	0.26	-0.34	1.6
		(1.48)	(0)	(0.24)	(0.3)	(0.14)	(0.16)	(0.14)	(0.14)	(0.12)	(0.42)
	glmpath	7	5	-	-	0.04	0.24	-0.21	0.22	-0.28	2.4
		(2.22)	(0)	-	-	(0.13)	(0.12)	(0.11)	(0.1)	(0.1)	(0.52)
	hybrid glmmlasso	6	5	0.89	0.87	0.08	1.05	-0.99	1	-1.03	0.44
		(1.48)	(0)	(0.43)	(0.58)	(0.19)	(0.25)	(0.23)	(0.18)	(0.16)	(0.32)
	hybrid glmpath	7	5	0.86	0.87	0.08	1.01	-0.99	0.99	-1.02	0.7
		(2.22)	(0)	(0.42)	(0.53)	(0.2)	(0.28)	(0.24)	(0.19)	(0.16)	(0.64)
	thres glmmlasso	10	5	1.02	1.11	0.1	1.19	-1.09	1.11	-1.13	1.3
		(3.71)	(0)	(0.7)	(0.85)	(0.22)	(0.29)	(0.23)	(0.2)	(0.19)	(0.77)
	thres glmpath	10	5	0.91	0.94	0.09	1.11	-1.07	1.11	-1.1	1.1
		(2.97)	(0)	(0.49)	(0.59)	(0.21)	(0.27)	(0.25)	(0.19)	(0.2)	(0.73)
H_2	oracle	5	5	0.89	0.94	0.11	1.02	-0.98	1.02	-1.02	0.13
		(0)	(0)	(0.4)	(0.53)	(0.18)	(0.25)	(0.15)	(0.18)	(0.16)	(0.1)
	glmmlasso	6	5	0.39	0.41	0.09	0.66	-0.31	0.27	-0.34	1.6
		(1.48)	(0)	(0.23)	(0.28)	(0.13)	(0.17)	(0.1)	(0.11)	(0.09)	(0.27)
	glmpath	6.5	5	-	-	0.08	0.23	-0.21	0.21	-0.28	2.4
		(0.74)	(0)	-	-	(0.11)	(0.13)	(0.08)	(0.11)	(0.08)	(0.34)
	hybrid glmmlasso	6	5	0.93	0.96	0.12	1.02	-0.99	1.05	-1.04	0.34
		(1.48)	(0)	(0.44)	(0.51)	(0.19)	(0.26)	(0.15)	(0.17)	(0.16)	(0.3)
	hybrid glmpath	6.5	5	0.87	0.94	0.12	1.01	-0.99	1.03	-1.04	0.48
		(0.74)	(0)	(0.42)	(0.5)	(0.18)	(0.22)	(0.15)	(0.18)	(0.17)	(0.37)
	thres glmmlasso	14	5	1.3	1.33	0.16	1.26	-1.16	1.2	-1.22	2
		(5.93)	(0)	(0.87)	(0.79)	(0.27)	(0.27)	(0.28)	(0.26)	(0.24)	(1.7)
	thres glmpath	13.5	5	0.9	1.03	0.17	1.17	-1.07	1.13	-1.15	1.8
		(5.19)	(0)	(0.52)	(0.64)	(0.24)	(0.25)	(0.19)	(0.22)	(0.21)	(1.2)

Table 2: *Simulation results (medians) for the logistic mixed models H_1 (rescaled median absolute deviations in parentheses). A * means that the corresponding coefficient is not subject to penalization in the GLMMLasso^{LA} estimate.*

Model	Method	$\lvert S(\hat{\beta})\rvert$	TP	$\hat{\theta}_1^2$	$\hat{\theta}_2^2$	$\hat{\rho}$	$\hat{\beta}_1^*$	$\hat{\beta}_2^*$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	SE
True		5	5	1	1	0.5	0.1	1	-1	1	-1	
H_1	oracle	5	5	0.88	0.94	0.53	0.1	0.97	-1.03	1.02	-1.01	0.14
		(0)	(0)	(0.46)	(0.54)	(0.37)	(0.18)	(0.24)	(0.17)	(0.15)	(0.15)	(0.1)
	glmmlasso	6	5	0.41	0.41	0.63	0.07	0.66	-0.33	0.28	-0.34	1.6
		(1.48)	(0)	(0.22)	(0.25)	(0.51)	(0.14)	(0.16)	(0.12)	(0.11)	(0.11)	(0.35)

ACCEPTED MANUSCRIPT

Table 3: *Results for the epilepsy data. Model III is based on 6 fixed-effect covariates while the other methods are based on* $p = 4000$ *variables, including 3994 noise covariates.* † *indicates that the corresponding coefficient is significant at the* 5% *level.* ‡ *means that five noise variables are selected, but not shown in the table.* $S(\hat{\boldsymbol{\beta}}) = \{k : \hat{\beta}_k \neq 0\}$ *is the total number of selected variables.*

	Model III	glmmlasso	hybrid glmmlasso	thres glmmlasso
BIC	527.3	571.8	515.5	480.3
$S(\hat{\boldsymbol{\beta}})$	6	2	2	$7^{\ddagger}$
Intercept	$1.58^{\dagger}$	1.62	1.58	1.58
Base	$0.66^{\dagger}$	$< 10^{-4}$	0.74	0.75
Trt	$-0.47^{\dagger}$	-	-	-
Base x Trt	0.36	-	-	-
Age	0.11	-	-	-
V4	-0.04	-	-	-
$\hat{\theta}^2_{sub}$	0.21	0.68	0.25	0.28
$\hat{\theta}^2_{obs}$	0.13	0.12	0.13	0.04

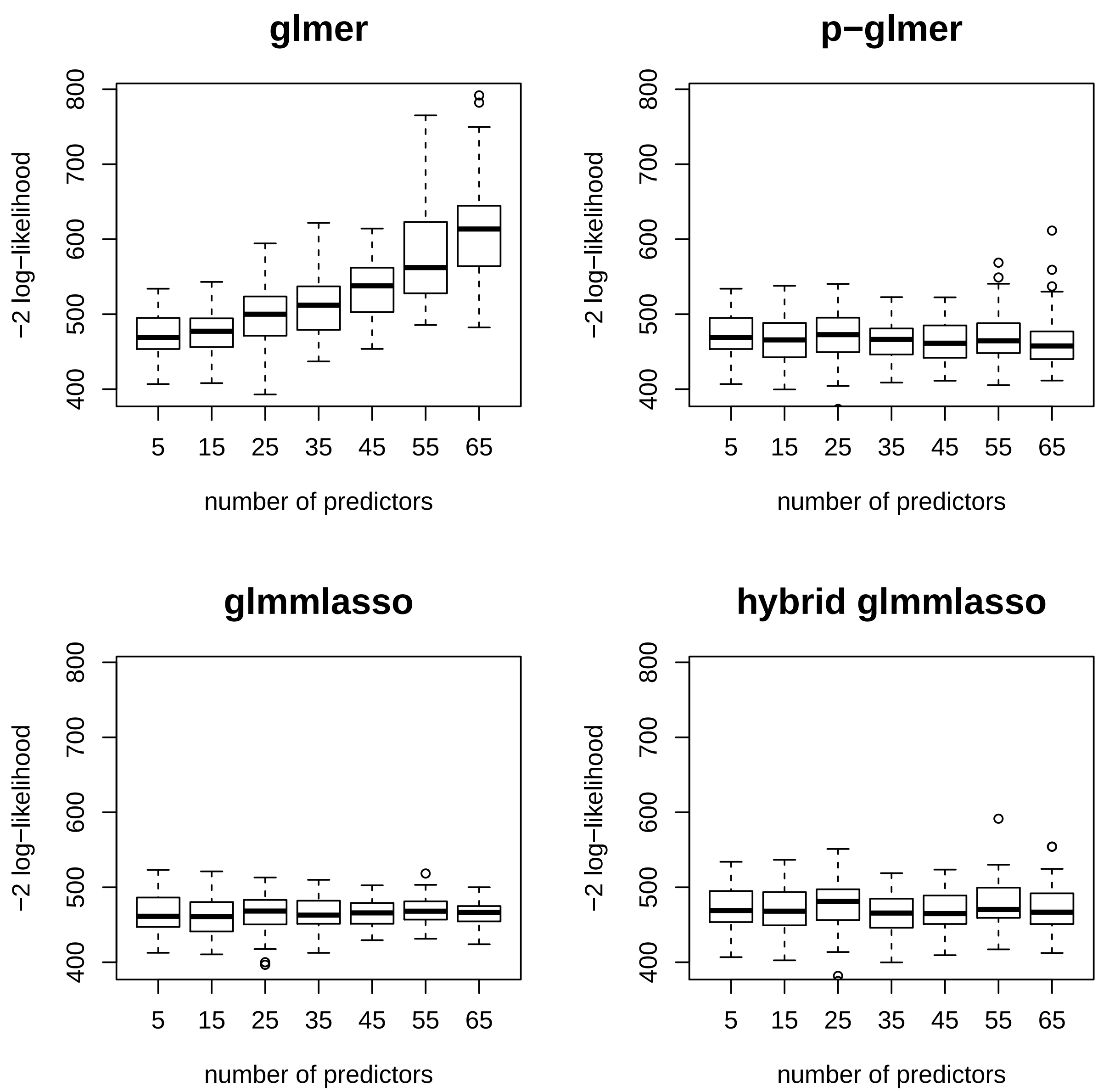

Figure 1: Minus twice out-of-sample log-likelihood for a growing number of covariates. The ML estimate performs badly whereas the GLMMLasso^{LA} estimators remain stable, and they are comparable to the p-glmer in the low-dimensional framework.

ACCEPTED MANUSCRIPT