Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ_1-Penalization

Jürg Schelldorfer and Peter Bühlmann
Seminar für Statistik
ETH Zürich
8092 Zürich

February 19, 2010

Abstract

We propose an ℓ_1-penalized estimation procedure for high-dimensional linear mixed-effects models. The models are useful whenever there is a grouping structure among high-dimensional observations, i.e. for clustered data. We prove a consistency and an oracle optimality result and we develop an algorithm with provable numerical convergence. Furthermore, we demonstrate the performance of the method on simulated and a real high-dimensional dataset.

1 INTRODUCTION

In recent years, high-dimensional linear regression problems have been extensively studied. The underlying assumption that all observations are independent is not always appropriate. We model high-dimensional data using a clustering structure such that observations between clusters are independent, but within each cluster they are dependent. This can be incorporated using mixed-effects models with high-dimensional, low sample size data. Mixed-effects models are an extension of linear models including random effects in addition to fixed effects. For example, many applications concern longitudinal data where the random effects vary between clusters and thereby induce a dependence structure within the clusters. It is a crucial and important question how to cope with high-dimensional mixed-effects models. Surprisingly, there is no established procedure for this problem which is well understood in terms of statistical properties. Difficulties arising from non-convexity of the likelihood may be a reason why high-dimensional linear mixed-effects models have not been approached so far. Besides methodology and theory, we will also empirically illustrate that there is a striking prediction improvement if we take into account the dependence structure in the data.

To deal with high-dimensionality, we suggest a Lasso-type procedure (Tibshirani, 1996). Assuming that the number of potential fixed effects is large and that the underlying true fixed-effects vector is sparse, we propose an ℓ_1-penalization on the fixed effects to achieve sparsity. Due to the presence of random effects, the maximum likelihood approach leads to the problem of a non-convex loss function whereas the main focus in high-dimensional computation and theory is devoted to convex loss functions. From an algorithmic point of view, we develop a coordinate gradient descent method and prove its numerical convergence to a stationary point. Regarding statistical properties, we establish consistency and an oracle optimality result.

Most of the existing literature in high-dimensional statistics deals with linear models. Greenshtein and Ritov (2004) prove that the Lasso consistently estimates the regression function under a sparsity condition in terms of the ℓ_1-norm of the regression parameter. More generally, van de Geer (2008) studies the prediction error for Lipschitz loss functions in the context of ℓ_1-penalized generalized linear models. Meinshausen and Bühlmann (2006) show that under the so-called neighborhood stability condition the Lasso does consistent variable selection. This condition is equivalent to the irrepresentable condition used in Zhao and Yu (2006). Both conditions are sufficient and (essentially) necessary for consistent model selection. Furthermore, assuming different conditions on the design, Bunea et al. (2007), van de Geer (2008), Zhang and Huang (2008), Meinshausen and Yu (2009), Bickel et al. (2009) study the behaviour of the Lasso for the estimation error between the estimated and true high-dimensional coefficient vectors. Similarly, Candes and Tao (2007) propose the Dantzig selector and derive estimation error bounds, and Bickel et al. (2009) show equivalent theoretical behaviour of the Lasso and the Dantzig selector.

Apart from the statistical properties, there has been a substantial interest in developing fast algorithms for solving problems with convex loss function and ℓ_1-type constraints. Efron et al. (2004) present the lars algorithm, but coordinate descent methods have outperformed the lars algorithm with respect to speed. Such optimization methods are used by Meier et al. (2008), Wu and Lange (2008) and Friedman et al. (2009). In fact, the entire regularization path can now be computed extremely fast even for datasets with thousands of covariates.

This paper is organised as follows. In Section 2 we define the ℓ_1-penalized linear mixed-effects model. In Section 3, we present the theoretical results for the ℓ_1-penalized estimator before describing the details of the algorithm in Section 4. After some simulations in Section 5 we apply the procedure to a real dataset. The technical proofs are deferred to the Appendix.

2 LINEAR MIXED-EFFECTS MODELS AND ℓ_1-PENALIZED ESTIMATION

In this article, we restrict ourselves to a relatively simple model. The estimation algorithm as well as the theory can be extended to more general linear mixed-effects models.

2.1 High-dimensional Model Set-up

We assume that the observations are inhomogeneous in the sense that they are not independent, but grouped. Let $i = 1, \dots, N$ be the grouping index and $j = 1, \dots, n_i$ the observation index within a group. Denote by $N_T = \sum_{i=1}^{N} n_i$ the total number of observations. For each observation, we observe a univariate response variable y_{ij} and a p-dimensional covariate $\boldsymbol{x}_{ij} \in \mathbb{R}^p$. We call $\boldsymbol{x}_{ij}$ the fixed-effects regression variables. Moreover, we have q-dimensional covariates $\boldsymbol{z}_{ij} \in \mathbb{R}^q$ which are called the random-effects regression variables.

We consider the following model:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{b} + \boldsymbol{z}_{ij}^T \boldsymbol{\beta}_i + \varepsilon_{ij} \qquad i = 1, \dots, N, \ \ j = 1, \dots, n_i \quad , \tag{1}$$

assuming that

i) $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and uncorrelated for $i = 1, \dots, N$ and $j = 1, \dots, n_i$,

ii) $(\boldsymbol{\beta}_i)_k \sim \mathcal{N}(0, \tau^2)$ and uncorrelated for $i = 1, \dots, N$ and $k = 1, \dots, q$,

iii) $\varepsilon_{11}, \dots, \varepsilon_{Nn_N}$ are independent of $\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_N$.

Here we denote by $\boldsymbol{b} \in \mathbb{R}^p$ the vector of the unknown fixed regression coefficients and by $\boldsymbol{\beta}_i \in \mathbb{R}^q$, $i = 1, \dots, N$, the random regression coefficients. As indicated by

the index i, the $\boldsymbol{\beta}_i$ are different among the groups. All observations have the coefficient $\boldsymbol{b}$ in common whereas the value of $\boldsymbol{\beta}_i$ depends on the group that the observation belongs to. In other words, for each group there are group-specific deviations $\boldsymbol{\beta}_i$ from the overall effects $\boldsymbol{b}$. We assume throughout the paper that the design variables $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are deterministic, i.e. fixed design.
Using standard notation in mixed-effects models (Pinheiro and Bates, 2000), we rewrite model (1) in the following notation:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{b} + \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad i = 1, \ldots, N \quad , \tag{2}$$

where $\boldsymbol{y}_i$ is a $n_i \times 1$ vector of responses of the ith group, $\boldsymbol{X}_i$ is a $n_i \times p$ fixed-effects design matrix, $\boldsymbol{b}$ is a $p \times 1$ vector of fixed regression coefficients, $\boldsymbol{Z}_i$ is a $n_i \times q$ random-effects design matrix, $\boldsymbol{\beta}_i$ is a vector of random regression coefficients with $\boldsymbol{\beta}_i \sim \mathcal{N}_q(\boldsymbol{0}, \tau^2\boldsymbol{I}_q)$ and $\boldsymbol{\varepsilon}_i$ is a $n_i \times 1$ error term with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}_{n_i}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$. As mentioned above, the design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are assumed to be deterministic. We allow that the number p of fixed-effects regression coefficients may be much larger than the total number of observations, i.e. $N_T \ll p$. Furthermore, the number q of random-effects variables might be as large as $q \leq p$. Nevertheless, we confine ourselves to keep q small, i.e. not larger than the number of observations per group. We aim at estimating the fixed regression parameter vector $\boldsymbol{b}$, the random effects $\boldsymbol{\beta}_i$ and the variance parameters σ^2 and τ^2. From model (2) we deduce that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ are independent with the following distributions: $\boldsymbol{y}_i \sim \mathcal{N}_{n_i}(\boldsymbol{X}_i\boldsymbol{b}, \boldsymbol{\Lambda}_i(\sigma^2, \tau^2))$ with $\boldsymbol{\Lambda}_i(\sigma^2, \tau^2) = \sigma^2\boldsymbol{I}_{n_i} + \tau^2\boldsymbol{Z}_i\boldsymbol{Z}_i^T$. Hence the negative log-likelihood function of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ is given by

$$-\ell(\boldsymbol{b}, \sigma^2, \tau^2) = \frac{1}{2}\sum_{i=1}^{N}\left\{n_i \log(2\pi) + \log|\boldsymbol{\Lambda}_i| + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})^T\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})\right\} \quad . \tag{3}$$

2.2 ℓ_1-penalized Maximum Likelihood Estimator

Due to the possibly large number of covariates ($N_T \ll p$ setting), we cannot use the classical maximum likelihood or restricted maximum likelihood approach. Assume that the fixed regression coefficients are sparse in the sense that many parameters are zero. We then attenuate these difficulties by adding an ℓ_1-penalty on the fixed regression coefficients. By doing so, we achieve a sparse solution with respect to the fixed effects. This leads us to consider the following objective function:

$$Q_\lambda(\boldsymbol{b}, \sigma^2, \tau^2) := \frac{1}{2}\sum_{i=1}^{N}\left\{\log|\boldsymbol{\Lambda}_i| + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})^T\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})\right\} + \lambda\sum_{k=2}^{p}|b_k| \quad , \tag{4}$$

where b_1 is the unpenalized intercept and λ a nonnegative regularization parameter. Consequently, we estimate the fixed regression coefficient vector $\boldsymbol{b}$ and the variance parameters σ^2 and τ^2 by

$$\hat{\boldsymbol{b}}, \hat{\sigma}^2, \hat{\tau}^2 = \underset{\boldsymbol{b}, \sigma^2, \tau^2}{\arg\min}\, Q_\lambda(\boldsymbol{b}, \sigma^2, \tau^2) \quad . \tag{5}$$

For fixed variance parameters σ^2, τ^2, the minimization with respect to $\boldsymbol{b}$ is a convex optimization problem. However, over all parameters, we have a non-convex objective function and hence, we have to deal with a non-convex problem. This requires a more general framework in theory as well as in computation. In the following Sections, we show how to address this issue.
Before, let us make some comments concerning the objective function (4): Since we want to make use of the convexity of $Q_\lambda(.)$ with respect to $\boldsymbol{b}$ (see Section 4), we do not profile the likelihood function in (3), as usually done in the mixed-effects model framework (Pinheiro and Bates, 2000), and the objective function (4) is non-convex only with respect to the variance parameters.

2.3 Prediction of the random-effects coefficients

We predict the random-effects coefficients $\boldsymbol{\beta}_i$, $i = 1, \ldots, N$ by the maximum a posteriori (MAP) principle. Denoting by f the density of the corresponding Gaussian random variable, we define

$$\begin{aligned}\boldsymbol{\beta}_i^* &= \arg\max_{\boldsymbol{\beta}_i} f(\boldsymbol{\beta}_i|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N, \boldsymbol{b}, \sigma^2, \tau^2) = \arg\max_{\boldsymbol{\beta}_i} f(\boldsymbol{\beta}_i|\boldsymbol{y}_i, \boldsymbol{b}, \sigma^2, \tau^2) \\ &= \arg\max_{\boldsymbol{\beta}_i} \frac{f(\boldsymbol{y}_i|\boldsymbol{\beta}_i) \cdot f(\boldsymbol{\beta}_i)}{f(\boldsymbol{y}_i)} = \arg\min_{\boldsymbol{\beta}_i} \left\{ \frac{1}{2}\frac{1}{\sigma^2}\|\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b} - \boldsymbol{Z}_i\boldsymbol{\beta}_i\|^2 + \frac{1}{2\tau^2}\|\boldsymbol{\beta}_i\|^2 \right\} \quad .\end{aligned}$$

From this we get $\boldsymbol{\beta}_i^* = [\boldsymbol{Z}_i^T\boldsymbol{Z}_i + \sigma^2/\tau^2\boldsymbol{I}_q]^{-1}\boldsymbol{Z}_i^T\boldsymbol{r}_i$ where $\boldsymbol{r}_i = (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})$ is the (marginal) residual vector. This solution corresponds to the well-known Ridge Regression with regularization parameter $\frac{\sigma^2}{\tau^2}$. Since the true values of $\boldsymbol{b}$, σ^2 and τ^2 are unknown, the $\boldsymbol{\beta}_i$'s are predicted by $\hat{\boldsymbol{\beta}}_i = [\boldsymbol{Z}_i^T\boldsymbol{Z}_i + \hat{\sigma}^2/\hat{\tau}^2\boldsymbol{I}_q]^{-1}\boldsymbol{Z}_i^T\hat{\boldsymbol{r}}_i$ with $\hat{\boldsymbol{r}}_i = (\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{b}})$, using the estimates from (5).

2.4 Selection of the regularization parameter

Estimation and selection of the fixed-effects coefficients require to choose the optimal regularization parameter λ. We propose to use the Bayesian Information Criterion (BIC) defined by

$$-2\ell(\hat{\boldsymbol{b}}, \hat{\sigma}^2, \hat{\tau}^2) + \log N_T \cdot \hat{df} \quad , \tag{6}$$

where $\hat{df} = |\{1 \leq j \leq p; \hat{b}_j \neq 0\}|$ is the number of the nonzero fixed regression coefficients. The use of $\hat{df}$ as a measure of the degrees of freedom is motivated by the work of Zou et al. (2007) who show that the expected number of degrees of freedom for the Lasso in a linear model is given by the number of nonzero coefficients.
Obviously, there are other tuning parameter selection methods, for example cross-validation and AIC-type criteria, among others. Advocating the BIC as selection criterion is based on our experience that it performs best in both simulations and real data examples (see Section 5).

2.5 Adaptive ℓ_1-penalized Maximum Likelihood Estimator

Due to the bias of the Lasso, Zou (2006) proposed the adaptive Lasso. The bias problem occurs more severely in the mixed-effects model setting than in linear regression. To be more specific, let us assume that the penalized kth covariate has a fixed- and a random-effects coefficient, i.e. b_k and $(\boldsymbol{\beta}_i)_k$, respectively. If λ is too large, $\hat{b}_k$ is shrunken too strongly towards zero. Thereby, the estimate of the variance parameter $\hat{\tau}^2$ is too large and also $(\hat{\boldsymbol{\beta}}_i)_k$ gets a bias related to the amount of shrinkage of $\hat{b}_k$.
To overcome this problem, we suggest employing an adaptive procedure. The adaptive ℓ_1-penalized maximum likelihood estimator uses the following objective function instead of (4):

$$Q_\lambda^{w_2,\ldots,w_p}(\boldsymbol{b}, \sigma^2, \tau^2) := \frac{1}{2}\sum_{i=1}^{N}\left\{ \log(|\boldsymbol{\Lambda}_i|) + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})^T\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b}) \right\} + \lambda\sum_{k=2}^{p} w_k|b_k|$$

and hence

$$\hat{\boldsymbol{b}}, \hat{\sigma}^2, \hat{\tau}^2 = \arg\min_{\boldsymbol{b},\sigma^2,\tau^2} Q_\lambda^{w_2,\ldots,w_p}(\boldsymbol{b}, \sigma^2, \tau^2) \tag{7}$$

It is vital to use an adaptive ℓ_1-penalty if there is at least one penalized variable having both a fixed and a random effect. The weights $w_2, \ldots, w_p$ can be derived from an initial estimation in (5) with $w_k = 1/|\hat{b}_{init,k}(\lambda)|$ for $k = 2, \ldots, p$. In the empirical examples of this paper, we employ the simple weights $w_k = 1/|\hat{b}_{init,k}(\lambda = 0+)|$ for $k = 2, \ldots, p$ where we choose $\lambda = 0+$ sufficiently small such that many variables are selected.

3 CONSISTENCY AND ORACLE INEQUALITY

In the high-dimensional setting with $p \gg N_T$, the theory for penalized estimation based on convex loss functions with an ℓ_1-penalty is well studied, see for example van de Geer (2008). From (4) and (5) we see that we are dealing with a non-convex loss function, due to the parameters σ^2 and τ^2, and a convex ℓ_1-penalty. We build here upon the theory for non-convex ℓ_1-penalized smooth likelihood problems, presented in Städler et al. (2009).

We use the following framework and notation. Let $i = 1, \ldots, N$ as before and $n_i \equiv n$ the same for all i. Denote by $\boldsymbol{y}_i \in \mathcal{Y} \subset \mathbb{R}^n$ the response variable. Let $\boldsymbol{X}_i$ be the fixed covariates in some space $\mathcal{X}^n \subset \mathbb{R}^{n \times p}$ and $\boldsymbol{Z}_i \subset \boldsymbol{X}_i$. The latter can be assumed without loss of generality, since we can assign to every variable a fixed effect being equal to zero. Define the parameter $\boldsymbol{\theta}^T := (\boldsymbol{b}^T, 2\log\sigma, 2\log\tau) = (\boldsymbol{b}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{p+2}$ and denote by $\boldsymbol{\theta}_0$ the true parameter vector. For the remaining part of the paper, we are using this parametrization. Furthermore, for a constant $0 < K < \infty$, consider the parameter space to be

$$\boldsymbol{\Theta} \subset \{\boldsymbol{\theta}^T = (\boldsymbol{b}^T, \boldsymbol{\eta}^T); \sup_{\boldsymbol{x} \in \mathcal{X}} |\boldsymbol{x}^T \boldsymbol{b}| \le K, \|\boldsymbol{\eta}\|_\infty \le K\} \in \mathbb{R}^{p+2} \quad , \tag{8}$$

where $\|\boldsymbol{\eta}\|_\infty = \max_{l=1,2} |\eta_l|$. We modify the estimator in (5) by restricting the solution to lie in the compact parameter space $\boldsymbol{\Theta}$:

$$\hat{\boldsymbol{\theta}}_\lambda := \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \Big\{ \log(|\boldsymbol{\Lambda}_i|) + (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{b})^T \boldsymbol{\Lambda}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{b}) \Big\} + \lambda \sum_{k=2}^{p} |b_k| \right\} \quad . \tag{9}$$

Now, let $\{f_{\boldsymbol{\theta}, \boldsymbol{X}_i, \boldsymbol{Z}_i}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be the Gaussian density for $\boldsymbol{y}_i$ with respect to the above parametrization. Since we use the negative log-likelihood as loss function, the excess risk coincides with the Kullback-Leibler distance:

$$\mathcal{E}_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \int \log\left(\frac{f_{\boldsymbol{\theta}_0, \boldsymbol{X}, \boldsymbol{Z}}}{f_{\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}}} \right) f_{\boldsymbol{\theta}_0, \boldsymbol{X}, \boldsymbol{Z}} \, d\mu \quad , \tag{10}$$

and we define the average excess risk as

$$\overline{\mathcal{E}}_{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N}(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{E}_{\boldsymbol{X}_i, \boldsymbol{Z}_i}(\boldsymbol{\theta}|\boldsymbol{\theta}_0) \quad .$$

In the sequel, we drop the indices ${}_{\boldsymbol{X},\boldsymbol{Z}}$ and ${}_{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N,\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_N}$, respectively.

3.1 Consistency

We require two conditions for consistency. The first assumption concerns the design matrix of the fixed-effects regression matrices $\boldsymbol{X}_i$. Write $\boldsymbol{X}_i^T = (\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_n^i)$ and let $\varsigma_{max}^2(.)$ be the largest eigenvalue of a square matrix.

Assumption 1 *The largest eigenvalue of*

$$\boldsymbol{\Sigma}_{N,n} := \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\boldsymbol{x}_j^i(\boldsymbol{x}_j^i)^T \qquad \in \mathbb{R}^{p\times p}$$

is bounded, i.e. there exist a constant L which does not depend on N, n and p such that $\varsigma^2_{max}(\Sigma_{N,n}) \leq L < \infty$.

The second assumption is a condition on the random-effects design matrices $\boldsymbol{Z}_i$.

Assumption 2 *Let $\big(\omega_j^{(i)}\big)_{j=1}^n$ be the eigenvalues of $\boldsymbol{Z}_i\boldsymbol{Z}_i^T$ for $i = 1,\ldots,N$.*

(a) The $\big(\omega_j^{(i)}\big)_{j=1}^n$ are bounded: $\omega_j^{(i)} \leq K < \infty$ for all i and j, with K from (8).

(b) At least two eigenvalues are different, i.e. for all i $\exists j_1 \neq j_2 \in \{1,\ldots,n\}$ such that $\omega_{j_1}^{(i)} \neq \omega_{j_2}^{(i)}$.

We consider a triangular scheme (Greenshtein and Ritov, 2004) of observations from (2):

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{b}_N + \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad i = 1,\ldots,N \quad , \tag{11}$$

where the parameters $\boldsymbol{b}_N$ and $\boldsymbol{\eta}_N$ are allowed to depend on N. We study consistency as $N \to \infty$ but the group size n is fixed. Moreover, let us use the notation $a \vee b := \max\{a, b\}$.

Theorem 1 - *Consistency.* *Consider model (11) and the estimator (9). Under Assumptions 1 and 2 (a) and assuming*

$$\|\boldsymbol{b}_{0,N}\|_1 = o\left(\sqrt{\frac{N}{\log^4(N)\log(p\vee N)}}\right), \quad \lambda_N = C\sqrt{\frac{\log^4(N)\log(p\vee N)}{N}} \quad \textit{for some } C > 0 \quad ,$$

any global minimizer $\hat{\boldsymbol{\theta}}_{\lambda_N}$ as in (9) satisfies $\overline{\mathcal{E}}(\hat{\boldsymbol{\theta}}_{\lambda_N}|\boldsymbol{\theta}_0) = o_P(1)$ for $N \to \infty$.

A proof is given in the Appendix. The condition on $\|\boldsymbol{b}_{0,N}\|_1$ is a sparsity condition on the true underlying fixed-effects coefficients.

3.2 Oracle Inequality

We now present an oracle optimality result in non-asymptotic form. Preliminary, we introduce some notation and present another assumption. Let $S(\boldsymbol{b}) = \{1 \leq j \leq p | b_j \neq 0\}$ be the active set of $\boldsymbol{b}$, i.e. the set of non-zero coefficients, and $\boldsymbol{b}_J = \{b_j | j \in J\}$ for $J \subset \{1,\ldots,p\}$. We denote by $S_0 = S(\boldsymbol{b}_0)$ the true active set and by $s_0 = |S_0|$ its cardinality.

Assumption 3 - *Restricted Eigenvalue Condition.* *There exists a constant $\kappa \geq 1$, such that for all $\boldsymbol{b} \in \mathbb{R}^p$ satisfying $\|\boldsymbol{b}_{S_0^c}\|_1 \leq 6\|\boldsymbol{b}_{S_0}\|_1$ it holds that $\|\boldsymbol{b}_{S_0}\|_2^2 \leq \kappa^2\boldsymbol{b}^T\Sigma_{N,n}\boldsymbol{b}$.*

A discussion of this assumption can be found in Bickel et al. (2009) and van de Geer and Bühlmann (2009). Define

$$\lambda_0 = M_N \log N\sqrt{\frac{\log(p\vee N)}{N}} \quad , \tag{12}$$

where M_N is of order $\log N$ and an exact definition is given in the proof of Theorem 1. For any $T \geq 1$, let $\mathcal{J}$ be a set defined by the underlying empirical process. It is shown in the proof of Theorem 1 that the set $\mathscr{J}$ has large probability,

$$\mathbb{P}[\mathscr{J}] \geq 1 - a_1 \exp\Big[- \frac{T^2 \log^2 N \log(p \vee N)}{a_2^2}\Big] - \frac{\rho}{\log N} \frac{1}{N^{1-2\varepsilon}}$$

for N sufficiently large and some constants $a_1, a_2, \varepsilon, \rho > 0$, see Lemma 2 and 3 in the Appendix.

Theorem 2 - *Oracle result.* *Consider the estimator (9). Under Assumptions 2 and 3, and for $\lambda \geq 2T\lambda_0$, then, on $\mathscr{J}$, for the average excess risk,*

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}}_\lambda | \boldsymbol{\theta}_0) + 2(\lambda - T\lambda_0) \|\hat{\boldsymbol{b}}_{S_0^c}\|_1 \leq 8(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0$$

for a constant c_0 (which is independent of N, n, p and the design).

A proof is given in the Appendix. Since $(\lambda - T\lambda_0) > 0$, we deduce from Theorem 2 that $\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}}_\lambda | \boldsymbol{\theta}_0) \leq 8(\lambda + T\lambda_0)^2 c_0^2 \kappa^2 s_0$. This means that the average Kullback-Leibler risk is of order $O(s_0 \lambda_0^2) = O(s_0 \frac{\log^4 N \log(p \vee N)}{N})$. Thus, up to a factor $\log^4 N \log(p \vee N)$, the optimal convergence rate if one knew the s_0 non-zero coefficients, is achieved. Moreover, because $\bar{\mathcal{E}}(\hat{\boldsymbol{\theta}}_\lambda | \boldsymbol{\theta}_0) \geq 0$, we conclude that $\|\hat{\boldsymbol{b}}_{S_0^c}\|_1 \leq 4(\lambda + T\lambda_0) c_0^2 \kappa^2 s_0$. This says that $\|\hat{\boldsymbol{b}}_{S_0^c}\|_1 = O(s_0 \lambda_0)$ and therefore, the noise variables from S_0^c have small estimated values.

4 COMPUTATIONAL ALGORITHM

The estimation of the regression parameter $\boldsymbol{b}$ and the variance parameters σ^2 and τ^2 is based on the Block Coordinate Gradient Descent (BCGD) method from Tseng and Yun (2009).

The main ideas of our BCGD algorithm are that we cycle through the coordinates and minimize the objective function $Q_\lambda(.)$ with respect to only one coordinate while keeping the other parameters fixed (i.e. a Gauss-Seidel algorithm). In each such step, we approximate $Q_\lambda(.)$ by a strictly convex quadratic function. Then, we calculate a descent direction and we employ an inexact line search to ensure a decrease in the objective function.

BCGD algorithms are used in Meier et al. (2008) for the grouped Lasso as well as in Wu and Lange (2008) and Friedman et al. (2009) for the ordinary Lasso. We remark that Meier et al. (2008) have a block structure due to the grouped variables whereas we only focus on ungrouped covariates. Thus the word "block" has no meaning in our context and consequently, we omit it in the subsequent discussion. Furthermore, the ordinary Lasso has only regression parameters to cycle through in contrast to our problem involving two kinds of parameters: fixed regression and variance parameters.

First, we introduce the notation and give an overview of the algorithm. Second, we focus on the details as well as on computational issues. And third, we show that our optimization problem achieves numerical convergence.

Let $\boldsymbol{\theta}^T = (\boldsymbol{b}^T, \boldsymbol{\eta}) \in \mathbb{R}^{p+2}$ be the parametrization introduced in the previous Section. Define the functions

$$P(\boldsymbol{\theta}) := \sum_{k=2}^{p} |b_k| \quad \text{and} \quad g(\boldsymbol{\theta}) := \frac{1}{2} \sum_{i=1}^{N} \Big\{ \log|\boldsymbol{\Lambda}_i(\boldsymbol{\eta})| + (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{b})^T \boldsymbol{\Lambda}(\boldsymbol{\eta})_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{b}) \Big\} \quad .$$

Then, (9) can be written as $\hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_\lambda(\boldsymbol{\theta}) := g(\boldsymbol{\theta}) + \lambda P(\boldsymbol{\theta})$.

Moreover, let $\mathcal{I}(\boldsymbol{\theta})$ be the Fisher information of $g(\boldsymbol{\theta})$ and $\boldsymbol{e}_j$ be the jth unit vector. For $k = 1, 2, 3, \ldots$, let $\mathcal{S}^k$ be the index cycling through the coordinates $\{1\}$,

$\{2\}$,..., $\{p\}$, $\{p+1\}$, $\{p+2\}$. Then, the computational algorithm can be summarized in the following way:

Algorithm 1 - ***Coordinate Gradient Descent.***

Step 0. Let $\boldsymbol{\theta}^0 \in \mathbb{R}^{p+2}$ *be an initial value.*

For $k = 0, 1, 2, \ldots$, *let* $\mathcal{S}^k$ *be the index cycling through the coordinates* $\{1\}$, $\{2\}$,..., $\{p\}$, $\{p+1\}$, $\{p+2\}$.

Step 1. Choose an approximate Hessian $H^k > 0$.

Step 2. $d^k := \arg\min_d \left\{ g(\boldsymbol{\theta}^k) + \frac{\partial}{\partial \theta_{\mathcal{S}_k}} g(\boldsymbol{\theta}^k) d + 1/2 d^2 H^k + \lambda P(\boldsymbol{\theta}^k + d\boldsymbol{e}_{\mathcal{S}_k}) \right\}$.

Step 3. Choose a stepsize $\alpha^k > 0$ *by the Armijo rule and set* $\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \alpha^k d^k \boldsymbol{e}_{\mathcal{S}_k}$.

The step length α^k is chosen in such a way that in each step, there is an improvement in the objective function $Q_\lambda(.)$. The Armijo rule is defined as follows:

Armijo Rule: *Choose* $\alpha^k_{init} > 0$ *and let* α^k *be the largest element of* $\{\alpha^k_{init}\delta^l\}_{l=0,1,2,..}$ *satisfying*

$$Q_\lambda(\boldsymbol{\theta}^k + \alpha^k d^k \boldsymbol{e}_{\mathcal{S}_k}) \leq Q_\lambda(\boldsymbol{\theta}^k) + \alpha^k \varrho \triangle^k \tag{13}$$

where $\triangle^k := \partial/\partial\theta_{\mathcal{S}_k} g(\boldsymbol{\theta}^k) d^k + \gamma (d^k)^2 H^k + \lambda P(\boldsymbol{\theta}^k + d^k \boldsymbol{e}_{\mathcal{S}_k}) - \lambda P(\boldsymbol{\theta}^k)$.
The choice of the constants comply with the suggestions in Bertsekas (1999) and are $\delta = 0.1, \varrho = 0.001, \gamma = 0$ and $\alpha^k_{init} = 1$.

We now turn to some details of Algorithm 1.

Initial value $\boldsymbol{\theta}^0$: As a starting value, we choose an ordinary Lasso solution by cross-validation ignoring the grouping structure among the observations. By doing so, we ensure that we are at least as good (with respect to the objective function) as an ordinary Lasso in a linear model.

Choice of H^k: For numerical convergence (see Theorem 3 below), we require that H^k is positive definite and bounded. Hence we use the Fisher information $\mathcal{I}(\boldsymbol{\theta})$ and, as proposed in Tseng and Yun (2009), for constants c_{min} and c_{max} we set $H^k = \min(\max(\mathcal{I}(\boldsymbol{\theta})_{\mathcal{S}_k\mathcal{S}_k}, c_{min}), c_{max})$.

Calculation of d^k: We have to distinguish whether the index $\mathcal{S}^k$ appears in $P(\boldsymbol{\theta})$ or not:

$$d^k = \begin{cases} \text{median}\left(\dfrac{\lambda - \frac{\partial}{\partial\theta_{\mathcal{S}_k}} g(\boldsymbol{\theta}^k)}{H^k}, -b_{\mathcal{S}_k}, \dfrac{-\lambda - \frac{\partial}{\partial\theta_{\mathcal{S}_k}} g(\boldsymbol{\theta}^k)}{H^k} \right) & \mathcal{S}^k \in \{2, \ldots, p\}, \\ -\frac{\partial}{\partial\theta_{\mathcal{S}_k}} g(\boldsymbol{\theta}^k)/H^k & \mathcal{S}^k \in \{1, p+1, p+2\}. \end{cases} \tag{14}$$

If $H^k = \mathcal{I}(\boldsymbol{\theta})_{\mathcal{S}_k\mathcal{S}_k}$, we take advantage of the fact that $g(\boldsymbol{\theta})$ is quadratic with respect to $\boldsymbol{b}$. Using $\alpha^k_{init} = 1$, the stepsize α^k chosen by the Armijo rule ($l = 0$) leads to the minimum of $g(\boldsymbol{\theta}^k)$ with respect to $b_{\mathcal{S}_k}$. The update $\hat{b}^{k+1}_{\mathcal{S}_k}(\lambda)$ is then given analytically by

$$\hat{b}^{k+1}_{\mathcal{S}_k}(\lambda) = \text{sign}\left(\sum_{i=1}^N (\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i) \boldsymbol{\Lambda}_i^{-1} \boldsymbol{x}^{(i)}_{\mathcal{S}_k} \right) \frac{\left(|\sum_{i=1}^N (\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i) \boldsymbol{\Lambda}_i^{-1} \boldsymbol{x}^{(i)}_{\mathcal{S}_k}| - \lambda \right)_+}{\sum_{i=1}^N \boldsymbol{x}^{(i)T}_{\mathcal{S}_k} \boldsymbol{\Lambda}_i^{-1} \boldsymbol{x}^{(i)}_{\mathcal{S}_k}}, \tag{15}$$

where $\boldsymbol{X}_i = (\boldsymbol{x}^{(i)}_1, \ldots, \boldsymbol{x}^{(i)}_p)$, $\tilde{\boldsymbol{y}}_i = \boldsymbol{X}_i^{(-\mathcal{S}_k)} \hat{\boldsymbol{b}}^k_{(-\mathcal{S}_k)}$ (leaving out the $\mathcal{S}_k$th variable), $(.)_+ = \max(., 0)$ and sign(.) the signum function.

Choice of the λ-sequence: We choose a λ_1 sufficiently large such that all penalized coefficients are zero. We calculate a sequence $\lambda_1 > \lambda_2 > \ldots$ on a log-scale until a model with a certain sparsity level is reached. At latest, we stop if the number of selected fixed-effects variables is larger than the total number of observations.

Active-Set Algorithm: Assuming that the solution is sparse, we can reduce the computing time by using an active-set algorithm, which is used in Meier et al. (2008) and Friedman et al. (2009). More specifically, we do not cycle through all coordinates, but we restrict ourselves to the current active set $S(\hat{\boldsymbol{b}})$ and update all coordinates of $\hat{\boldsymbol{b}}$ only every Dth iteration. This reduces the computational time remarkably.

Theorem 3 - ***Convergence of the CGD algorithm.*** *If $(\boldsymbol{\theta}^k)_{k\geq 0}$ is chosen according to Algorithm 1, then every cluster point of $\{\boldsymbol{\theta}^k\}_{k\geq 0}$ is a stationary point of $Q_\lambda(\boldsymbol{\theta})$.*

Proof. It remains to check that the assumptions in Tseng and Yun (2009) are fulfilled. More precisely: $\lambda > 0$, $P(.) = |.|_1$ is a proper, convex, continuous function and blockseparable with respect to $\mathcal{S}^k$, $g(.)$ is continuously differentiable on $dom(P) = \{\boldsymbol{\theta}|P(\boldsymbol{\theta}) < \infty\}$, $c_{min} \leq H^k \leq c_{max}$ for $k \geq 1$ and $0 < c_{min} \leq c_{max},$. Moreover, $\sup_k \alpha^k > 0$ and $\inf_k \alpha^k_{init} > 0$. □

Due to the non-convexity of the optimization problem, our CGD algorithm is not finding a global optimum. However, it finds a global optimum for fixed σ^2 and τ^2 parameters and hence, a global optimum over all the parameters can be found by applying our CGD algorithm on a grid for (σ^2, τ^2).

5 SIMULATION STUDY AND REAL DATA APPLICATION

In this section, we asses the empirical performance of the adaptive ℓ_1-penalized maximum likelihood estimator (7) in different kinds of examples. We study several performance measures and make a comparison to other ℓ_1-penalization procedures before illustrating the method on a real dataset.

5.1 Simulation Studies

We will focus on the following characteristics in a series of simulation examples. We study the variable selection performance and we investigate the estimation accuracy for the variance and the fixed regression parameters. More specifically, due to the proposed penalized maximum likelihood approach, we investigate the bias of the estimated parameters. Moreover, we will differentiate between three types of fixed-effects regression parameters. The intercept b_1 which is not subject to ℓ_1-penalization and for $b_k, k = 2, \ldots, p$, we distinguish if the kth variable has a random effect $(\boldsymbol{\beta}_i)_k$ or not.

In all subsequent simulation schemes, we restrict ourselves to the case where all groups have the same number of observations, i.e. we set $n \equiv n_i$ for $i = 1, \ldots, N$. We assign $\boldsymbol{Z}_i \subset \boldsymbol{X}_i$ such that the columns of $\boldsymbol{Z}_i$ correspond to the first q columns of $\boldsymbol{X}_i$. This means that the first q variables have both a fixed-effects coefficient b_k and a random-effects coefficient $(\boldsymbol{\beta}_i)_k$ for $i = 1, \ldots, N$ and $k = 1, \ldots, q$. For $i = 1, \ldots, N$ and $j = 1, \ldots, n$ the covariates are generated according to $(\boldsymbol{x}_{ij})_{(-1)} \sim \mathcal{N}_{p-1}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with the pairwise correlation $\boldsymbol{\Sigma}_{ll'} = \rho^{|l-l'|}$ for $l, l' = 2, \ldots, p$ $(l, l' = 1$ is the intercept). In the following, $\boldsymbol{b} = (b_1, ..., b_p)$ comprises the non-penalized intercept b_1 and the penalized coefficients $b_2, \ldots, b_p$ and we denote by $s_0 := \#\{1 \leq j \leq p; (b_0)_j \neq 0\}$ the true number of non-zero coefficients. We also report the signal-to-noise ratio (SNR). It is defined by

$$SNR := \frac{\mathrm{Var}\,(\boldsymbol{b}^T \boldsymbol{x}_{ij} + \boldsymbol{\beta}^T \boldsymbol{z}_{ij})}{\mathrm{Var}\,(\varepsilon)} = \frac{q\tau^2 + \boldsymbol{b}_{(-1)}^T \boldsymbol{\Sigma} \boldsymbol{b}_{(-1)}}{\sigma^2} \quad . \tag{16}$$

Firstly, we give an example in the low-dimensional setting. The design is chosen similar to examples presented in Pinheiro and Bates (2000) and the choice of the variance parameters to those in Jiang and Rao (2003).

M1: $N = 10$, $p = 10$, $n = 7$, $q = 3$, $\sigma = 1$, $\tau = 1$ and $s_0 = 4$ with $\boldsymbol{b}_0 = (1, 1, 2, 1, 0, \ldots, 0)^T$.

Secondly, we study three example in the high-dimensional setting.

M2: $N = 15$, $p = 300$, $n = 5$, $q = 4$, $\sigma = 1$, $\tau = 1$ and $s_0 = 6$ with $\boldsymbol{b}_0 = (1, 1, 2, 3, 1, 1, 0, \ldots, 0)^T$.

M3: $N = 30$, $p = 500$, $n = 6$, $q = 1$, $\sigma = 1$, $\tau = 1$ and $s_0 = 4$ with $\boldsymbol{b}_0 = (1, 1, 2, 3, 0, \ldots, 0)^T$.

M4: $N = 12$, $p = 1000$, $n = 6$, $q = 2$, $\sigma = 1$, $\tau = 1$ and $s_0 = 4$ with $\boldsymbol{b}_0 = (1, 1.5, 1, 1, 0, \ldots, 0)^T$.

We always use the BIC criterion (6) to choose the regularization parameter λ. The results in the form of means and standard deviations (in parentheses) over 100 simulation runs are reported in Table 1. Therein, $|S(\hat{\boldsymbol{b}})|$ denotes the cardinality of the estimated active set, TP is the number of true positives, $\hat{b}_1$ denotes the non-penalized intercept, $\hat{b}_2$ is the coefficient of the first penalized variable having both a fixed and random regression coefficient and $\hat{b}_{q+1}$ is the first penalized covariate which only has a fixed regression coefficient.

Let us now summarize the results concerning variable selection. The estimated average active set is sparse and only slightly larger than the cardinality of the true active set $S_0 = S(\boldsymbol{b}_0)$. This property might be expected because it is known from linear regression that on the one hand the BIC selects a sparse model and on the other hand the adaptive Lasso remarkably reduces the number of false positives (FP) compared to the ordinary Lasso. The results of the number of true positives (TP) indicate that the algorithm is mainly able to identify the true fixed-effects coefficients. As can be seen from model **M3**, the algorithm mainly misses true fixed effects if there are variables with both a fixed- and a random-effects coefficient. Although employing an adaptive approach, we are not able to overcome this problem completely.

To sum up the parameter estimation accuracy, we see that the variance parameter estimates are biased towards zero. We observe that a maximum likelihood approach (in contrast to a restricted maximum likelihood approach) gives biased variance estimators.

In all models, the (unpenalized) intercept is estimated well although it has a random-effects coefficient $(\boldsymbol{\beta}_i)_1$ for $i = 1, \ldots, N$. For the (penalized) coefficients, we see a marked difference between the columns $\hat{b}_2$ and $\hat{b}_{q+1}$ concerning both means and standard deviations. In contrast to the $(q+1)$th variable, the 2nd variable has a random-effects coefficient $(\boldsymbol{\beta}_i)_2$. Table 1 suggests that the presence of a random-effects coefficient increases the bias of the fixed-effects coefficient as well as the estimation accuracy. To conclude, the reduction of the bias of the ℓ_1-penalized method (not shown) is feasible for coefficients without a random effect by using an

Table 1: Simulation results for the adaptive ℓ_1-penalized maximum likelihood estimator (7).

Model	ρ	SNR	$\lvert S(\hat{\boldsymbol{b}})\rvert$	TP	$\hat{\sigma}$	$\hat{\tau}$	$\hat{b}_1$	$\hat{b}_2$	$\hat{b}_{q+1}$
M1	0.2	10.68	4.28	3.80	0.96	0.94	1.01	0.76	0.95
($s_0 = 4$)			(1.10)	(0.43)	(0.11)	(0.18)	(0.38)	(0.46)	(0.17)
	0.5	13.50	4.09	3.72	0.99	0.95	1.04	0.71	0.95
			(0.85)	(0.45)	(0.12)	(0.19)	(0.36)	(0.51)	(0.19)
	0.8	16.68	4.32	3.63	0.96	0.95	0.96	0.77	0.94
			(1.16)	(0.54)	(0.13)	(0.18)	(0.33)	(0.59)	(0.38)
M2	0.2	25.49	11.04	5.72	0.78	0.88	1.03	0.60	0.89
($s_0 = 6$)			(6.13)	(0.51)	(0.23)	(0.21)	(0.33)	(0.47)	(0.27)
	0.5	36.88	9.43	5.69	0.83	0.91	1.01	0.62	0.98
			(5.81)	(0.51)	(0.21)	(0.23)	(0.33)	(0.47)	(0.34)
	0.8	53.33	9.11	5.42	0.84	0.90	1.03	0.72	1.05
			(5.72)	(0.67)	(0.20)	(0.19)	(0.29)	(0.55)	(0.62)
M3	0.2	18.44	5.82	4	0.97	0.96	0.98	-	0.96
($s_0 = 4$)			(8.49)	(0)	(0.11)	(0.16)	(0.19)	-	(0.09)
	0.5	24.50	5.10	4	0.96	0.96	1.04	-	0.97
			(2.49)	(0)	(0.07)	(0.13)	(0.22)	-	(0.11)
	0.8	31.64	4.79	4	0.98	0.95	1.01	-	0.92
			(1.87)	(0)	(0.06)	(0.16)	(0.21)	-	(0.15)
M4	0.2	5.88	9.97	3.72	0.81	0.85	1.01	0.64	0.92
($s_0 = 4$)			(9.90)	(0.45)	(0.24)	(0.28)	(0.38)	(0.48)	(0.20)
	0.5	7.5	9.39	3.72	0.80	0.89	0.98	0.68	0.99
			(8.47)	(0.47)	(0.24)	(0.28)	(0.35)	(0.52)	(0.22)
	0.8	9.48	9.51	3.77	0.82	0.81	0.99	0.86	1.00
			(8.28)	(0.45)	(0.23)	(0.25)	(0.35)	(0.48)	(0.43)

NOTE: Means and standard deviations (in parentheses) for the cardinality of the active set $\lvert S(\hat{\boldsymbol{b}})\rvert$, the number of true positives TP, the variance parameter estimations $\hat{\sigma}$ and $\hat{\tau}$ and some selected fixed regression coefficients ($\hat{b}_1$ is the non-penalized intercept ; $\hat{b}_2$ is the first penalized covariate with both a fixed- and random-effects coefficient (”-” indicates that there is no such covariate) ; $\hat{b}_{q+1}$ is the first penalized covariate with no additional random-effects coefficient).

adaptive Lasso procedure. In contrast, good identification and estimation of variables with fixed- and random-effects parameters remains challenging in ℓ_1-penalized approaches.

We now turn to consider the performance of the proposed methodology concerning prediction. We compare the predictive performance with the Lasso and the adaptive Lasso, which both do not consider a grouping structure. In detail, we make a comparison between four different Lasso procedures. In doing so, denote by LME-Lasso the adaptive estimator in (7). As before, the best model is determined by minimizing the BIC on a grid of λ-values. Then denote by BIC-Lasso the lars algorithm (i.e. the Lasso) which evaluates the BIC at each transition point of the regularization path. Additionally, we use a cv-Lasso and a cv-adaptive Lasso whose optimal λ-value is chosen by 10-fold cross-validation. We fix the following scenario: $N = 25$, $n_i \equiv 6$ for $i = 1, \ldots, N$, $q = 3$, $s_0 = 5$ with $\boldsymbol{b}_0 = (1, 1.5, 1.2, 1, 2, 0, \ldots, 0)^T$, $\sigma = 1$ and $\rho = 0.2$. We only alter the number of fixed covariates p and the variance parameter τ^2. For measuring the quality of prediction, we generate a validation set with 50 observations per group and calculate the mean squared prediction error. The three models considered are

$$\textbf{M5}: p = 10, \textbf{M6}: p = 100 \text{ and } \textbf{M7}: p = 500.$$

The results are shown in Table 2.

Table 2: Means squared prediction error for three simulation examples.

Model	τ^2	LME-Lasso	BIC-Lasso	cv-Lasso	cv-adaptive Lasso
M5	0	1.02	1.00	1.06	1.02
(p=10)	0.25	1.35	1.76	1.88	1.92
	1	1.63	3.74	3.78	3.66
	2	1.74	5.92	5.85	6.16
M6	0	1.13	1.26	1.24	1.19
(p=100)	0.25	1.38	1.75	1.99	1.67
	1	2.25	4.35	4.60	4.22
	2	2.12	7.04	7.13	7.04
M7	0	1.08	2.13	1.33	1.66
(p=500)	0.25	1.62	3.58	2.84	3.47
	1	2.03	7.97	4.11	4.82
	2	2.17	15.59	9.44	11.27

We see that the methods differ slightly for $\tau^2 = 0$ which corresponds to no grouping structure. As τ^2 increases, the mean squared prediction error rises less for the LME-Lasso than for the other three Lasso methods. These results highlight that we can indeed achieve prediction improvements using the suggested mixed-effects model approach if the underlying model is given by (2).

5.2 Application: Riboflavin Data

We illustrate the proposed procedure on a real data set which is provided by DSM (Switzerland). The response variable is the logarithm of the riboflavin production rate of Bacillus subtilis. There are $p = 4088$ covariates measuring the gene expression levels. We take $N = 28$ samples (groups) with $n_i \in \{2, \ldots, 6\}$ and $N_T = 111$ observations.

Preliminary, we address the issue of determining those covariates which have both a fixed and a random regression coefficient. In other words, we have to find the matrix $\boldsymbol{Z}_i \subset \boldsymbol{X}_i$. Fitting first a model with $q = p$ reveals that it is reasonable to fit a so called random-intercept model wherein only the intercept has an additional random effect. This model can be written as

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{b} + \beta_{i1} + \varepsilon_{ij} \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i \tag{17}$$

The resulting variance estimates obtained from the ℓ_1-penalized (unweighted) estimator (5) are $\hat{\sigma} = 0.67$ and $\hat{\tau} = 0.22$. This means that 10% of the total variation is explained by the variation of the intercept between the N groups. The estimated active set $|S(\hat{\boldsymbol{b}})|$ comprises nine covariates. As might have been expected from the simulation studies, the size of the active set is smaller than that of the Lasso (18 variables) and that of the adaptive Lasso (12 variables). Nevertheless, the variables with the largest absolute value coincide in all three methods.

In the next step, we focus on the predictive performance of the ℓ_1-penalized estimator (5) in model (17) compared to the ordinary Lasso. By doing so, we reduce the data further such that $n_i = 4$ for all the remaining samples. This easily allows for conducting a leave-one-timepoint out cross-validation. As a result, the mean squared prediction error of the estimator (5) in model (17) is about 13% smaller than that of the ordinary Lasso.

6 DISCUSSION

We present an ℓ_1-penalized maximum likelihood estimator for high-dimensional linear mixed-effects models. The proposed methodology copes with the difficulty of combining a non-convex loss function and an ℓ_1-penalty. Thereby, we deal with theoretical and computational aspects which are substantially more challenging than in the linear regression setting. We prove theoretical results concerning the consistency of the estimator and we present a non-asymptotic oracle result. Moreover, by developing a coordinate gradient descent algorithm, we achieve provable numerical convergence of our algorithm to at least a stationary point. Our simulation studies and real data example show that the predictive performance can be remarkably improved when incorporating the knowledge about the cluster structure among observations.

APPENDIX: TWO TECHNICAL PROOFS

We present here the proofs of the theorems in Section 3.

A.1 Proof of Theorem 1

The proof consists of three parts. Firstly, we need an inequality ensuring that Lemma 2 holds. Secondly, we show that the probability (A.2) in Lemma 2 is large. And for completion of our proof, we can then refer to Städler et al. (2009).
From model (2), the log-likelihood function of $\boldsymbol{y}_i$ with respect to the parametrization in (8) is given by

$$\ell_{\boldsymbol{\theta}}(\boldsymbol{y}_i) := -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|e^{\eta_2}\boldsymbol{Z}_i\boldsymbol{Z}_i^T + e^{\eta_1}\boldsymbol{I}| - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})^T(e^{\eta_2}\boldsymbol{Z}_i\boldsymbol{Z}_i^T + e^{\eta_1}\boldsymbol{I})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{b})$$

Then, define the score function $s_{\boldsymbol{\theta}}(\boldsymbol{y}_i) := \partial/\partial\boldsymbol{\theta}\ell_{\boldsymbol{\theta}}(\boldsymbol{y}_i)$.

Lemma 1 *Under Assumption 2 (a), there exist constants* $c_1, c_2, c_3 \in \mathbb{R}_+$ *such that*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|s_{\boldsymbol{\theta}}(\boldsymbol{y}_i)\|_\infty \leq G_1(\boldsymbol{y}_i) := c_1 + c_2\|\boldsymbol{y}_i\|_2 + c_3\|\boldsymbol{y}_i\|_2^2 \quad i = 1, \ldots, N \quad .$$

Proof. The proof is straightforward using the Cauchy-Schwarz inequality and the fact that the induced L_2-norm of a square, symmetric matrix $\boldsymbol{A}$ is given by $\|\boldsymbol{A}\|_2 = \sqrt{\mathrm{tr}\,(\boldsymbol{A}^2)}$, where $\mathrm{tr}(\boldsymbol{A})$ denotes the trace of $\boldsymbol{A}$. □

Now we introduce the empirical process and present a result which controls the increments of it. The Lemma below gives a lower bound for the probability that the increments are small. Afterwards, we show that this lower bound is large.

Define the empirical process

$$V_N(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N}\Big\{\ell_{\boldsymbol{\theta}}(\boldsymbol{y}_i) - \mathbb{E}[\ell_{\boldsymbol{\theta}}(\boldsymbol{y}_i)]\Big\}$$

and

$$\lambda_0 = M_N \log N \sqrt{\frac{\log(p \vee N)}{N}} \quad . \tag{A.1}$$

Lemma 2 *Assume Assumption 1 and 2 (a). For constants a_1, a_2 and a_3 depending on L and K and for all $T \geq 1$,*

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\Big| V_N(\boldsymbol{\theta}) - V_N(\boldsymbol{\theta}_0) \Big|}{(\|\boldsymbol{b} - \boldsymbol{b}_0\|_1 + \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\|_2) \vee \lambda_0} \leq a_3 T \lambda_0$$

with probability at least

$$1 - a_1 \exp\Big[- \frac{T^2 \log^2 N \log(p \vee N)}{a_2^2} \Big] - \mathbb{P}\Bigg(\frac{1}{N} \sum_{i=1}^{N} F(\boldsymbol{y}_i) > \frac{T\lambda_0^2}{dK} \Bigg) \tag{A.2}$$

where $d := n + 2$ and

$$F(\boldsymbol{y}_i) = G_1(\boldsymbol{y}_i)\mathbf{1}_{\{G_1(\boldsymbol{y}_i) > M_N\}} + \mathbb{E}\Big[G_1(\boldsymbol{y}_i)\mathbf{1}_{\{G_1(\boldsymbol{y}_i) > M_N\}} \Big] \quad . \tag{A.3}$$

The proof of Lemma 2 is given in Städler et al. (2009). Next, we show that the third term is small in our setting.

Lemma 3 *There are constants b_1 and b_2 depending on K and n, a constant ρ depending on T, n and K such that for any $0 < \varepsilon < 1/2$ and $M_N := b_1(2\sqrt{\log N} + \sqrt{b_2})^2$ we have*

$$\mathbb{P}\Bigg(\frac{1}{N} \sum_{i=1}^{N} F(\boldsymbol{y}_i) > \frac{T\lambda_0^2}{dK} \Bigg) \leq \frac{\rho}{\log N} \frac{1}{N^{1-2\varepsilon}} \quad .$$

Proof. In the subsequent discussion, if A is a constant, we assume throughout that N is large enough such that $M_N - A > 0$. From (A.1) we see that it suffices to show that for a constant a_4,

$$\mathbb{P}\Bigg(\frac{1}{N} \sum_{i=1}^{N} F(\boldsymbol{y}_i) > a_4 \frac{\log N}{N} \Bigg) \leq \frac{\rho}{\log N} \frac{1}{N^{1-2\varepsilon}} \quad . \tag{A.4}$$

The expectation in (A.3) only affects the constants in the remainder of the proof. Therefore, we omit this term in the sequel. From

$$\mathbb{P}[c_1 + c_2\|\boldsymbol{y}_i\|_2 + c_3\|\boldsymbol{y}_i\|_2^2 > M_N] \leq \mathbb{P}\Big[\|\boldsymbol{y}_i\|_2^2 > \Big(\frac{M_N - c_1}{2c_2} \Big)^2 \Big] + \mathbb{P}\Big[\|\boldsymbol{y}_i\|_2^2 > \frac{M_N - c_1}{2c_3} \Big] \quad ,$$

and the fact that $M_N \to \infty$, we deduce that we can reduce ourselves to the analysis of $\mathbb{P}[\|\boldsymbol{y}_i\|_2^2 > M_N]$. For the sake of notational simplicity, we will leave out the index i and show that for an appropriate definition of M_N,

$$\mathbb{P}[\|\boldsymbol{y}\|_2^2 > M_N] \leq \frac{n}{N^2} \quad . \tag{A.5}$$

Denote by $\chi_\nu^2(\delta)$ the noncentral χ^2 distribution with ν degrees of freedom and noncentrality parameter δ. The following identity holds (Liu et al., 2008).

Claim 1 *If $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$ positive definite, then $\|\boldsymbol{y}\|_2^2 = \boldsymbol{y}^T\boldsymbol{y} = \sum_{j=1}^{n} \lambda_j \chi_1^2(\delta_j)$ where $\{\chi_1^2(\delta_j)\}_{j=1}^{n}$ are independent, λ_j, $j = 1, \ldots, n$ are the eigenvalues of $\boldsymbol{\Lambda}$ and if $\boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ for an orthonormal matrix $\boldsymbol{U}$, then $\delta_j = (\boldsymbol{U}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\mu})_j^2$.*

Claim 2

$$\mathbb{P}[\chi_1^2(\delta) > M] \leq \frac{1}{\sqrt{M}-\sqrt{\delta}} \frac{2}{\sqrt{2\pi}} \exp\left(- \frac{(\sqrt{M}-\sqrt{\delta})^2}{2} \right) \quad .$$

Proof. If $X \sim \mathcal{N}(\mu, \zeta^2)$, then by definition of the noncentral χ^2 distribution: $(X/\zeta)^2 \sim \chi^2_{\nu=1}(\delta = (\mu/\zeta)^2)$. Hence

$$\mathbb{P}[\chi_1^2(\delta) > M] = 2 \cdot \mathbb{P}[\frac{X}{\zeta} > \sqrt{M}] = 2 \cdot \mathbb{P}[\frac{X-\mu}{\zeta} > \sqrt{M} - \sqrt{\delta}] = 2 \cdot S(\sqrt{M}-\sqrt{\delta}) \quad , \tag{A.6}$$

where $S(t) := \frac{1}{\sqrt{2\pi}} \int_t^\infty \exp(-u^2/2) du$ is the survival function of a standard Gaussian random variable for which the following inequalities holds:

$$\frac{t}{1+t^2} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) < S(t) < \frac{1}{t} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \quad \text{for} \quad t > 0 \quad .$$

Thus, we conclude

$$\mathbb{P}[\chi_1^2(\delta) > M] \leq \frac{1}{\sqrt{M}-\sqrt{\delta}} \frac{2}{\sqrt{2\pi}} \exp\left(- \frac{(\sqrt{M}-\sqrt{\delta})^2}{2} \right) \quad .$$

□

Claim 3 *For* $M_{N,\delta} := (2\sqrt{\log N} + \sqrt{\delta})^2$*,*

$$\mathbb{P}[\chi_1^2(\delta) > M_{N,\delta}] \leq \frac{1}{N^2}.$$

Proof. Using Claim 2,

$$\begin{aligned}\mathbb{P}[\chi_1^2(\delta) > M_{N,\delta}] &\leq \frac{1}{\sqrt{M_{N,\delta}}-\sqrt{\delta}} \frac{2}{\sqrt{2\pi}} \exp(-\frac{(\sqrt{M_{N,\delta}}-\sqrt{\delta})^2}{2}) \\ &\leq 1 \cdot \exp(-\frac{(2\sqrt{\log N}+\sqrt{\delta}-\sqrt{\delta})^2}{2}) \leq \frac{1}{N^2}.\end{aligned}$$

□

Claim 4 *For the eigenvalues* $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_n)$ *of* $\boldsymbol{\Lambda}$*,* $\boldsymbol{\delta} := (\delta_1, \ldots, \delta_n)$*,* $M_{N,n,\boldsymbol{\lambda},\delta_j} := n\lambda_{max}(2\sqrt{\log N} + \sqrt{\delta_j})^2$ *(*λ_{max} *is the maximal eigenvalue of* $\boldsymbol{\Lambda}$*), define* $\delta := \arg\max_{\delta_j, 1\leq j \leq n} \mathbb{P}[\chi_1^2(\delta_j) > \frac{M_{N,n,\boldsymbol{\lambda},\delta_j}}{n\lambda_{max}}]$ *and set* $M_{N,n,\boldsymbol{\lambda},\boldsymbol{\delta}} = M_{N,n,\boldsymbol{\lambda},\delta}$*, then*

$$\mathbb{P}[\|\boldsymbol{y}\|_2^2 > M_{N,n,\boldsymbol{\delta},\boldsymbol{\lambda}}] \leq \frac{n}{N^2}$$

Proof. For any $M > 0$, using Claim 1 and 2

$$\begin{aligned}\mathbb{P}[\|\boldsymbol{y}\|_2^2 > M] &= \mathbb{P}[\sum_{j=1}^n \lambda_j \chi_1^2(\delta_j) > M] \leq \sum_{j=1}^n \mathbb{P}[\chi_1^2(\delta_j) > \frac{M}{n\lambda_j}] \leq \sum_{j=1}^n \mathbb{P}[\chi_1^2(\delta_j) > \frac{M}{n\lambda_{max}}] \\ &\leq n \cdot \max_{1\leq j\leq n} \mathbb{P}[\chi_1^2(\delta_j) > \frac{M}{n\lambda_{max}}]\end{aligned}$$

Set $M = M_{N,n,\boldsymbol{\lambda},\boldsymbol{\delta}}$ and using Claim 3:

$$\mathbb{P}[\|\boldsymbol{y}\|_2^2 > M_{N,n,\boldsymbol{\lambda},\boldsymbol{\delta}}] \leq n \cdot \mathbb{P}[\chi_1^2(\delta) > (2\sqrt{\log N} + \sqrt{\delta})^2] \leq \frac{n}{N^2} \quad .$$

□

At this point, we have proven (A.5). We now use this result to derive formula (A.4). Due to Assumption 2 (a), $\lambda_{max}^{(i)} \leq e^K(1+K^2) := b_1$ and $\delta_j^{(i)} \leq nKe^K := b_2$ for all i and j. Thereby, we use the value $M_N = b_1(2\sqrt{\log N} + \sqrt{b_2})^2$. Moreover, we use the Markov inequality and the Hölder inequality for any $0 < \varepsilon < 1/2$.

$$
\begin{aligned}
&\mathbb{P}\Bigg[\frac{1}{N}\sum_{i=1}^{N} G_1(\boldsymbol{y}_i)\mathbf{1}_{\{G_1(\boldsymbol{y}_i)>M_N\}} > a_4\frac{\log N}{N}\Bigg] \\
&= \mathbb{P}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\Big[c_1 + c_2\|\boldsymbol{y}_i\| + c_3\|\boldsymbol{y}_i\|_2^2\Big]\mathbf{1}_{\{c_1+c_2\|\boldsymbol{y}_i\|+c_3\|\boldsymbol{y}_i\|_2^2>M_N\}} > a_4\frac{\log N}{N}\Bigg] \\
&\leq \frac{1}{a_4}\frac{1}{\log N}\Bigg\{c_1\sum_{i=1}^{N}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\|_2 + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big] + c_2\sum_{i=1}^{N}\mathbb{E}\Big[\|\boldsymbol{y}_i\|_2\mathbf{1}_{\{c_1+c_2\|\boldsymbol{y}_i\|_2+c_3\|\boldsymbol{y}_i\|_2^2>M_N\}}\Big] \\
&\quad + c_3\sum_{i=1}^{N}\mathbb{E}\Big[\|\boldsymbol{y}_i\|_2^2\mathbf{1}_{\{c_1+c_2\|\boldsymbol{y}_i\|+c_3\|\boldsymbol{y}_i\|_2^2>M_N\}}\Big]\Bigg\} \\
&\leq \frac{1}{a_4}\frac{1}{\log N}\Bigg\{c_1\sum_{i=1}^{N}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\|_2 + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big] \\
&\quad + c_2\mathbb{E}\Big[(\|\boldsymbol{y}_i\|_2)^{\frac{1}{\varepsilon}}\Big]^{\varepsilon}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\| + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big]^{1-\varepsilon} \\
&\quad + c_3\sum_{i=1}^{N}\mathbb{E}\Big[(\|\boldsymbol{y}_i\|_2^2)^{\frac{1}{\varepsilon}}\Big]^{\varepsilon}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\|_2 + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big]^{1-\varepsilon}\Bigg\} \\
&\leq \frac{1}{a_4}\frac{1}{\log N}\Bigg\{c_1\sum_{i=1}^{N}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\|_2 + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big] + \tilde{c}_2\sum_{i=1}^{N}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\| + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big]^{1-\varepsilon} \\
&\quad + \tilde{c}_3\sum_{i=1}^{N}\mathbb{P}\Big[c_1 + c_2\|\boldsymbol{y}_i\| + c_3\|\boldsymbol{y}_i\|_2^2 > M_N\Big]^{1-\varepsilon}\Bigg\} \\
&\leq \frac{2}{a_4}\frac{1}{\log N}\Big\{c_1\sum_{i=1}^{N}\frac{n}{N^2} + \tilde{c}_2\sum_{i=1}^{N}\Big(\frac{n}{N^2}\Big)^{1-\varepsilon} + \tilde{c}_3\sum_{i=1}^{N}\Big(\frac{n}{N^2}\Big)^{1-\varepsilon}\Big\} \\
&\leq \frac{\rho}{\log N}\frac{1}{N^{1-2\varepsilon}} \quad ,
\end{aligned}
$$

where we used Claim 4. This ends the proof of Lemma 3. □

Now, we have shown that the probability (A.2) in Lemma 2 is large. Defining the set $\mathcal{J}$ by

$$
\mathcal{J} = \left\{\sup_{\boldsymbol{\theta}^T=(\boldsymbol{b}^T,\boldsymbol{\eta}^T)\in\Theta}\frac{\Big|V_N(\boldsymbol{\theta}) - V_N(\boldsymbol{\theta}_0)\Big|}{(\|\boldsymbol{b}-\boldsymbol{b}_0\|_1 + \|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_2)\vee\lambda_0} \leq a_3T\lambda_0\right\} \tag{A.7}
$$

means that $\mathcal{J}$ has large probability. The rest of the proof of Theorem 1 is as in Städler et al. (2009).

A.2 Proof of Theorem 2

It is sufficient to check Conditions $1-3$ in Städler et al. (2009). Subsequently, each of these is stated as a Lemma and again for simplicity, we drop the index i.

Let us introduce a slightly different parametrization, which coincides with that in Städler et al. (2009) and which simplifies the proofs below. For $\boldsymbol{x}_k \in \mathbb{R}^p, k =$

$1, \dots, n$, define $\boldsymbol{X}^T = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. Let

$$\begin{aligned}\boldsymbol{\phi}^T &= \boldsymbol{\phi}(\boldsymbol{X})^T = (\boldsymbol{x}_1^T\boldsymbol{b}, \dots, \boldsymbol{x}_n^T\boldsymbol{b}, 2\log\sigma, 2\log\tau) = ((\boldsymbol{X}\boldsymbol{b})^T, \boldsymbol{\eta}^T) = (\boldsymbol{\xi}(\boldsymbol{X})^T, \boldsymbol{\eta}^T)\\ &= (\boldsymbol{\xi}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{n+2}\end{aligned}$$

be the parameter vector with dimension $d := n+2$. By (8), the parameter space is bounded by the constant K: $\boldsymbol{\Phi} \subset \{\boldsymbol{\phi} \in \mathbb{R}^d : \|\boldsymbol{\phi}\|_\infty \le K\}$ where $\|\boldsymbol{\phi}\|_\infty := \max_{1\le j\le d}|\phi_j|$. Let $\{f_{\boldsymbol{\phi}}(\boldsymbol{y}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$ be the Gaussian density of $\boldsymbol{y}$ and $\ell_{\boldsymbol{\phi}}(\boldsymbol{y})$ its log-likelihood function. Moreover, let $\boldsymbol{\phi}_0$ be the true parameter vector.

Lemma 4 *Under Assumption 2 (a) holds*

$$\sup_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \max_{(j_1,j_2,j_3)\in\{1,\dots,d\}^3} \left| \frac{\partial^3}{\partial\phi_{j_1}\partial\phi_{j_2}\partial\phi_{j_3}} \ell_{\boldsymbol{\phi}}(\boldsymbol{y}) \right| \le G_3(\boldsymbol{y}) \quad ,$$

where

$$\sup_{\boldsymbol{X}\in\mathcal{X}^n} \int G_3(\boldsymbol{y}) f_{\boldsymbol{\phi}_0}(\boldsymbol{y}) d\mu(\boldsymbol{y}) \le C_3 < \infty \quad .$$

Proof. Set $G_3(\boldsymbol{y}) := d_1 + d_2\|\boldsymbol{y}\|_2 + d_3\|\boldsymbol{y}\|_2^2$ for appropriate constants $d_1, d_2, d_3 \in \mathbb{R}_+$. The proof makes use of the same techniques as the proof of Lemma 1 in the Appendix A.1. □

Let $\boldsymbol{A}$ be a symmetric and positive definite matrix. Denote by $\varsigma^2_{min}(\boldsymbol{A})$ its smallest eigenvalue, by $\mathrm{tr}(\boldsymbol{A})$ its trace and by $|\boldsymbol{A}|$ its determinant.

Lemma 5 *Under Assumption 2 (b), the Fisher information matrix $\mathcal{I}(\boldsymbol{\xi}(\boldsymbol{X}), \boldsymbol{\eta})$ is strictly positive definite, and in fact $\inf_{\boldsymbol{X}\in\mathcal{X}^n} \varsigma^2_{min}(\mathcal{I}(\boldsymbol{\xi}(\boldsymbol{X}), \boldsymbol{\eta})) > 0$.*

Proof. For $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{\xi}, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = e^{\eta_1}\boldsymbol{I} + e^{\eta_2}\boldsymbol{Z}\boldsymbol{Z}^T$, the Fisher information matrix is given by

$$\mathcal{I}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \begin{pmatrix} 2\boldsymbol{\Lambda}^{-1} & 0 & 0 \\ 0 & \frac{1}{2}e^{2\eta_1}\,\mathrm{tr}(\boldsymbol{\Lambda}^{-2}) & \frac{1}{2}e^{\eta_1+\eta_2}\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\Lambda}^{-1}) \\ 0 & \frac{1}{2}e^{\eta_1+\eta_2}\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\Lambda}^{-1}) & \frac{1}{2}e^{2\eta_2}\,\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T) \end{pmatrix} .$$

The upper left part of the matrix is given by $\boldsymbol{D}_1 := 2\boldsymbol{\Lambda}^{-1}$, which is positive definite. Hence it remains to prove that the lower right block matrix $\boldsymbol{D}_2$ is also positive definite. Let $(\omega_j)_{j=1}^n$ be the eigenvalues of $\boldsymbol{Z}\boldsymbol{Z}^T$.

$$\mathrm{tr}(\boldsymbol{\Lambda}^{-2}) = \sum_{j=1}^n \frac{1}{(e^{\eta_2}\omega_j + e^{\eta_1})^2} \quad , \quad \mathrm{tr}(\boldsymbol{\Lambda}^{-2}\boldsymbol{Z}\boldsymbol{Z}^T) = \sum_{j=1}^n \frac{\omega_j}{(e^{\eta_2}\omega_j + e^{\eta_1})^2} \quad ,$$

$$\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T) = \sum_{j=1}^n \frac{\omega_j^2}{(e^{\eta_2}\omega_j + e^{\eta_1})^2} \quad .$$

Then

$$\begin{aligned}|\boldsymbol{D}_2| &= \frac{1}{4}e^{2(\eta_1+\eta_2)}\Big[\mathrm{tr}(\boldsymbol{\Lambda}^{-2})\cdot\mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T) - \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{\Lambda}^{-1})^2\Big]\\ &= \frac{1}{4}e^{2(\eta_1+\eta_2)}\Big[\sum_j\sum_{j'} \frac{1}{(e^{\eta_2}\omega_j + e^{\eta_1})^2}\frac{\omega_{j'}^2}{(e^{\eta_2}\omega_{j'} + e^{\eta_1})^2} - \sum_j\sum_{j'} \frac{\omega_j\omega_{j'}}{(e^{\eta_2}\omega_j + e^{\eta_1})^2(e^{\eta_2}\omega_{j'} + e^{\eta_1})^2}\Big]\\ &= \frac{1}{4}e^{2(\eta_1+\eta_2)}\Big[\sum_j\sum_{j'} \frac{1}{(e^{\eta_2}\omega_j + e^{\eta_1})^2(e^{\eta_2}\omega_{j'} + e^{\eta_1})^2}\Big[\omega_{j'}^2 - \omega_j\omega_{j'}\Big]\Big]\\ &= \frac{1}{4}e^{2(\eta_1+\eta_2)}\Big[\sum_{j<j'} \upsilon_{jj'}[\omega_j^2 - \omega_{j'}^2]^2\Big] > 0 \quad .\end{aligned}$$

The last equality holds due to the identity $\upsilon_{jj'} = \upsilon_{j'j}$. □

Lemma 6 *Under Assumption 2 (b), for all $\epsilon > 0$, there exists an $\alpha_\epsilon > 0$, such that*

$$\inf_{\boldsymbol{X}\in\mathcal{X}^n}\ \inf_{\boldsymbol{\phi}\in\boldsymbol{\Phi},\|\boldsymbol{\phi}-\boldsymbol{\phi}_0\|_2>\epsilon}\mathcal{E}(\boldsymbol{\phi}(\boldsymbol{X})|\boldsymbol{\phi}_0(\boldsymbol{X}))\geq\alpha_\epsilon.$$

Proof. Let $\boldsymbol{\phi}^T = (\boldsymbol{\xi}^T, \boldsymbol{\eta}^T)$, $\boldsymbol{\phi}_0^T = (\boldsymbol{\xi}_0^T, \boldsymbol{\eta}_0^T)$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}_0$ the corresponding covariance matrices. Then $\log f_{\boldsymbol{\phi}_0}(\boldsymbol{y}) - \log f_{\boldsymbol{\phi}}(\boldsymbol{y}) = \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{1}{2}\log|\boldsymbol{\Lambda}_0| + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\xi})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{y}-\boldsymbol{\xi}) - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\xi}_0)^T\boldsymbol{\Lambda}_0^{-1}(\boldsymbol{y}-\boldsymbol{\xi}_0)$. Since $\mathcal{E}(\boldsymbol{\phi}|\boldsymbol{\phi}_0) := \mathbb{E}_{\boldsymbol{\phi}_0}\Big[\log f_{\boldsymbol{\phi}_0}(\boldsymbol{y}) - \log f_{\boldsymbol{\phi}}(\boldsymbol{y})\Big]$, it follows

$$\mathcal{E}(\boldsymbol{\phi}|\boldsymbol{\phi}_0) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Lambda}|}{|\boldsymbol{\Lambda}_0|} + \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_0) + (\boldsymbol{\xi}_0-\boldsymbol{\xi})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{\xi}_0-\boldsymbol{\xi}) - n\right]$$

By definition of the excess risk $\mathcal{E}(\boldsymbol{\phi}|\boldsymbol{\phi}_0) \geq 0$. Denote by $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and $\boldsymbol{\eta}_0 = ((\eta_0)_1, (\eta_0)_2)$, then we can detail:

$$\log\frac{|\boldsymbol{\Lambda}|}{|\boldsymbol{\Lambda}_0|} = -\sum_{j=1}^n \log\left(\frac{e^{(\eta_0)_2}\omega_j + e^{(\eta_0)_1}}{e^{\eta_2}\omega_j + e^{\eta_1}}\right) \quad , \quad \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}_0) = \sum_{j=1}^n \frac{e^{(\eta_0)_2}\omega_j + e^{(\eta_0)_1}}{e^{\eta_2}\omega_j + e^{\eta_1}}$$

Thus, we get

$$\mathcal{E}(\boldsymbol{\phi}|\boldsymbol{\phi}_0) = \frac{1}{2}\Big\{(\boldsymbol{\xi}_0-\boldsymbol{\xi})^T\boldsymbol{\Lambda}^{-1}(\boldsymbol{\xi}_0-\boldsymbol{\xi})\Big\} + \frac{1}{2}\sum_{j=1}^n\left\{\frac{e^{(\eta_0)_2}\omega_j + e^{(\eta_0)_1}}{e^{\eta_2}\omega_j + e^{\eta_1}} - \log\left(\frac{e^{(\eta_0)_2}\omega_j + e^{(\eta_0)_1}}{e^{\eta_2}\omega_j + e^{\eta_1}}\right) - 1\right\}$$

The first term is strictly positive if $\boldsymbol{\xi}_0 \neq \boldsymbol{\xi}$ and zero iff $\boldsymbol{\xi}_0 = \boldsymbol{\xi}$. The second term is a function of the form $u - \log(u) - 1 \geq 0$ for $u \geq 0$. The second term is only zero if all terms are exactly zero. Due to Assumption 2 (b), we get the claim. □

References

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

Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007), "Sparsity oracle inequalities for the Lasso," *Electronic Journal of Statistics*, 1, 169–194.

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation when p is much larger than n," *The Annals of Statistics*, 35, 2313–2351.

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. (2009), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Preprint, available at http://www.cnbc.cmu.edu/cns/papers/glmnet.pdf*, .

Greenshtein, E., and Ritov, Y. (2004), "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization," *Bernoulli*, 10, 971–988.

Jiang, J., and Rao, J. S. (2003), "Consistent Procedures for Mixed Linear Model Selection," *The Indian Journal of Statistics*, 65, 23–42.

Liu, H., Tang, Y., and Zhang, H. H. (2008), "A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables," *Computational Statistics and Data Analysis*, .

Meier, L., van de Geer, S., and Bühlmann, P. (2008), "The group lasso for logistic regression," *Journal of the Royal Statistical Society*, 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.

Meinshausen, N., and Yu, B. (2009), "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, 37, 246–270.

Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-Plus*, New York: Springer.

Städler, N., Bühlmann, P., and van de Geer, S. (2009), "l_1-Penalization for Mixture Regression Models," *To appear in Test (with discussion), available at ftp://ftp.stat.math.ethz.ch/Research-Reports/Other-Manuscripts/buhlmann/stadbuhlgeer-final.pdf*, .

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

Tseng, P., and Yun, S. (2009), "A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization," *Mathematical Programming, Series 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., and Bühlmann, P. (2009), "On the conditions used to prove oracle results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392.

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

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the lasso selection in high-dimensional linear regression," *The Annals of Statistics*, 36, 1567–1594.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learing Research 7*, pp. 2541–2563.

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.