Scandinavian Journal of STATISTICS

Theory and Applications

Scandinavian Journal of Statistics, Vol. 38: 197–214, 2011
doi: 10.1111/j.1467-9469.2011.00740.x

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

JÜRG SCHELLDORFER, PETER BÜHLMANN and SARA VAN DE GEER
Seminar für Statistik, ETH Zurich

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, that is, 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 data set.

Key words: adaptive Lasso, coordinate gradient descent, coordinatewise optimization, Lasso, random-effects model, variable selection, variance components

1. Introduction

1.1. High-dimensional statistical inference: some known results for convex loss functions

Substantial progress has been achieved over the last decade in high-dimensional statistical inference where the number of parameters p is allowed to be of much larger order than sample size n. To fix ideas, suppose we focus on estimation of a p-dimensional parameter $\boldsymbol{\beta}_0$ based on n noisy observations where $p \gg n$. Although such a problem is ill-posed in general, it can be accurately solved if the underlying true structure of $\boldsymbol{\beta}_0$ is sparse. Here, sparsity may be measured in terms of the ℓ_r-norm $\|\boldsymbol{\beta}\|_r = (\sum_{j=1}^p |\beta_j|^r)^{1/r}$ $(0 \le r < \infty)$. Very roughly speaking, high-dimensional statistical inference is possible, in the sense of leading to reasonable accuracy or asymptotic consistency, if

$$\log(p) \cdot \text{sparsity}(\boldsymbol{\beta}_0)^{\alpha} \ll n,$$

where typically $\alpha = 2$ [cf. (1)] or $\alpha = 1$ [cf. (2)], and assuming that the underlying (e.g. regression) design behaves reasonably.

A lot of attention has been devoted to high-dimensional linear models

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon},$$

with $n \times p$ design matrix $\mathbf{X}$ and $p \gg n$. A very popular and powerful estimation method is the Lasso, proposed by Tibshirani (1996). It is an acronym for least absolute shrinkage and selection operator and the name indicates that the method does some variable selection in the sense that some of the regression coefficient estimates are exactly zero. Among the main reasons why it has become very popular for high-dimensional estimation problems are its statistical accuracy for prediction and variable selection coupled with its computational feasibility which involves convex optimization only. The latter is in sharp contrast to exhaustive variable selection based on least squares estimation whose computational complexity is in general exponential in p. The statistical properties of the Lasso in high-dimensional settings have been worked out in numerous articles. Without (essentially) a condition on the design $\mathbf{X}$, the Lasso satisfies:

This paper is based on a plenary presentation at NORDSTAT Meeting in Voss Norway, June 2010.

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/n = O_P(\|\boldsymbol{\beta}_0\|_1 \sqrt{\log(p)/n}), \tag{1}$$

where $O_P(\cdot)$ is with respect to $p \geq n \to \infty$ (Bühlmann & van de Geer, 2011). That is, if the model is sparse with $\|\boldsymbol{\beta}_0\|_1 \ll \sqrt{\log(p)/n}$, we obtain consistency. Such kind of a result has been proved by Greenshtein & Ritov (2004). Later, optimality has been established where (1) is improved to

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/n = O_P(s_0 \xi^{-2} \log(p)/n),$$

and furthermore

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_r = O_P(s_0^{1/r} \xi^{-2} \sqrt{\log(p)/n}), \quad r \in \{1, 2\}, \tag{2}$$

where s_0 equals the number of non-zero coefficients and ξ^2 denotes a restricted eigenvalue of the design matrix $\boldsymbol{X}$ (Bühlmann & van de Geer, 2011). The rate in (2) is optimal up to the $\log(p)$ factor and the restricted eigenvalue ξ^2: oracle least squares estimation where the relevant variables would be known would have rate $O_P(s_0/n)$. We emphasize that for obtaining optimal convergence rates as in (2), we need to make some assumptions on the design that ξ^2 is not getting too small as $p \geq n \to \infty$, something we do not require in (1). Works dealing with various aspects around (2) include Bunea *et al.* (2007), van de Geer (2008), Zhang & Huang (2008), Meinshausen & Yu (2009) and Bickel *et al.* (2009).

A quite different problem is variable selection for inferring the true underlying active set $S_0 = \{1 \leq k \leq p : \beta_{0,k} \neq 0\}$. A simple estimator is $\hat{S} = \{1 \leq k \leq p : \hat{\beta}_k \neq 0\}$ where no significance testing is involved. Meinshausen & Bühlmann (2006) show for the Lasso that under the so called neighbourhood stability condition for the design, the Lasso does consistent variable selection in the sense that

$$\mathbb{P}[\hat{S} = S_0] \to 1 (p \geq n \to \infty), \tag{3}$$

assuming that the non-zero coefficients in S_0 are sufficiently large in absolute value, for example, $\min_{k \in S_0} |\beta_{0,k}| \gg \xi^{-2} \sqrt{s_0 \log(p)/n}$ which is the rate in (2) for $r = 2$. The neighbourhood stability condition is equivalent to the irrepresentable condition used in Zhao & Yu (2006), and they are both sufficient and (essentially) necessary for consistent model selection as in (3). Unfortunately, the neighbourhood stability and the irrepresentable condition are rather restrictive and many designs $\mathbf{X}$ would violate them. In case of (weaker) restrictive eigenvalue conditions, one still has the variable screening property for the Lasso

$$\mathbb{P}[\hat{S} \supseteq S_0] \to 1 \ (p \geq n \to \infty), \tag{4}$$

again assuming that the non-zero coefficients in S_0 are sufficiently large in absolute value. Equation (4) says that the Lasso does not miss a relevant variable from S_0; in addition, for the Lasso, the cardinality $|\hat{S}| \leq \min(n, p)$ and hence, for $p \gg n$, we achieve a huge dimensionality reduction in (4). The adaptive Lasso, proposed by Zou (2006) is a two-stage method which achieves (3) under a weaker restrictive eigenvalue assumption than the irrepresentable condition (Huang *et al.*, 2008; van de Geer *et al.*, 2010). We summarize the basic facts in Table 1. Moreover, everything essentially holds in an analogous way when using the Lasso in generalized linear models, that is, ℓ_1-norm penalization of the negative log-likelihood (van de Geer, 2008). Finally, we note that Bickel *et al.* (2009) prove equivalent theoretical behaviour of the Lasso and the Dantzig selector (Candes & Tao, 2007) in terms of (2), exemplifying that properties like (2) hold for other estimators than the Lasso as well.

Having some variable screening property as in (4), we can reduce the false positive selections by various methods, besides the adaptive Lasso mentioned above, including also stability selection (Meinshausen & Bühlmann, 2010) based on subsampling or via assigning p-values (Meinshausen *et al.*, 2009; Wasserman & Roeder, 2009) based on sample splitting.

Table 1. *Properties of the Lasso and required conditions to achieve them. Restricted eigenvalue assumption is weaker than the neighbourhood stability or irrepresentable condition* (*van de Geer & Bühlmann*, 2009). *For the adaptive Lasso*: *variable selection as in* (3) *can be achieved under restricted eigenvalue conditions*

Property	Design condition	Size of non-zero coefficients
Consistency as in (1)	No requirement	No requirement
Fast convergence rate as in (2)	Restricted eigenvalue	No requirement
Variable selection as in (3)	Neighbourhood stability $\Leftrightarrow$ irrepresentable condition	Sufficiently large
Variable screening as in (4)	Restricted eigenvalue	Sufficiently large

Regarding computation, the Lasso involves convex optimization. Popular algorithms are based on the homotopy method (Osborne *et al.*, 2000) such as LARS (Efron *et al.*, 2004). More recently, it has been argued that the coordinate gradient descent approach (in a Gauss–Seidel manner) is typically more efficient (Meier *et al.*, 2008; Wu & Lange, 2008; Friedman *et al.*, 2010).

1.2. High-dimensional linear mixed-effects models with non-convex loss function

The underlying assumption that all observations are independent is not always appropriate. We consider here linear mixed-effects models (Laird & Ware, 1982; Pinheiro & Bates, 2000; Verbeke & Molenberghs, 2000; Demidenko, 2004) where high-dimensional data incorporate a grouping structure with independent observations between and dependence within groups. Mixed-effects models, including random besides fixed effects, are a popular extension of linear models in that direction. For example, many applications concern longitudinal data where the random effects vary between groups and thereby induce a dependence structure within groups. It is a crucial and important question how to cope with high-dimensional linear mixed-effects models. Surprisingly, for this problem, there is no established procedure which is well understood in terms of statistical properties.

The main difficulty arises from non-convexity of the negative log-likelihood function which makes computation and theory very challenging. We are presenting some methodology, computation and theory for ℓ_1-norm penalized maximum likelihood estimation in linear mixed-effects models where the number of fixed effects may be much larger than the overall sample size but the number of covariance parameters of the random effects part being small. Based on a framework for ℓ_1-penalization of smooth but non-convex negative log-likelihood functions (Städler *et al.*, 2010), we develop in section 3 analogues of (1), (2) and (4), see also Table 1, and some properties of an adaptively ℓ_1-penalized estimator. In our view, these are the key properties in high-dimensional statistical inference in any kind of model. For example, with (4) at hand, p-values for single fixed-effects coefficients could be constructed along the lines of Meinshausen *et al.* (2009), controlling the familywise error or false discovery rate (but we do not apply such a method in this article). Furthermore, we design in section 4 an efficient coordinate gradient descent algorithm for linear mixed-effects models which is proved to converge numerically to a stationary point of the corresponding non-convex optimization problem.

We remark that we focus here on the case where it is prespecified which covariates are modelled with a random effect and which are not. In some situations, this is fairly realistic: for example, a random intercept model is quite popular and often leads to a reasonable model fit. Without prespecification of the covariates having a random effect, one could do variable selection based on penalized likelihood approaches on the level of random effects:

this has been developed from a methodological and computational perspective by Bondell *et al.* (2010) and Ibrahim *et al.* (2010) for low-dimensional settings. Addressing such problems in the truly high-dimensional scenario is beyond the scope of this article. However, we present in section 6 a real high-dimensional data problem where some exploratory analysis is used for deciding which covariates are to be modelled with a random effect. This example also illustrates empirically that there is a striking improvement if we incorporate random effects into the model, in comparison with a high-dimensional linear model fit.

The rest of this article is organized as follows. In section 2, we define the ℓ_1-penalized linear mixed-effects estimator. In section 3, we present the theoretical results for this estimator before describing the details of a computational algorithm in section 4. After some simulations in section 5 we apply the procedure to a real data set. The technical proofs are deferred to an Appendix in the online Supporting Information on the journal website.

2. Linear mixed-effects models and ℓ_1-penalized estimation

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 $i=1,\ldots,N$ be the grouping index and $j=1,\ldots,n_i$ the observation index within a group. Denote by $N_T=\sum_{i=1}^N n_i$ the total number of observations. For each group, we observe an $n_i\times 1$ vector of responses $\mathbf{y}_i$, and let $\mathbf{X}_i$ be an $n_i\times p$ fixed-effects design matrix, $\boldsymbol{\beta}$ a $p\times 1$ vector of fixed regression coefficients, $\mathbf{Z}_i$ an $n_i\times q$ random-effects design matrix and $\mathbf{b}_i$ a group-specific vector of random regression coefficients.

Using the notation from Pinheiro & Bates (2000), the model can be written as:

$$\mathbf{y}_i=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{b}_i+\boldsymbol{\varepsilon}_i \quad i=1,\ldots,N, \tag{5}$$

assuming that

(i) $\boldsymbol{\varepsilon}_i\sim\mathcal{N}_{n_i}(\mathbf{0},\sigma^2\mathbf{I}_{n_i})$ and uncorrelated for $i=1,\ldots,N$,
(ii) $\mathbf{b}_i\sim\mathcal{N}_q(\mathbf{0},\boldsymbol{\Psi})$ and uncorrelated for $i=1,\ldots,N$,
(iii) $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_N,\mathbf{b}_1,\ldots,\mathbf{b}_N$ are independent.

Here, $\boldsymbol{\Psi}=\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ is a general covariance matrix where $\boldsymbol{\theta}$ is an unconstrained set of parameters (with dimension q^*) such that $\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ is positive definite (i.e. by using the Cholesky decomposition). Possible structures for $\boldsymbol{\Psi}$ may be a multiple of the identity, a diagonal or a general positive definite matrix. We would like to remark that assumption (i) can be generalized to (i') $\boldsymbol{\varepsilon}_i\sim\mathcal{N}_{n_i}(\mathbf{0},\sigma^2\boldsymbol{\Lambda}_i)$ with $\boldsymbol{\Lambda}_i=\boldsymbol{\Lambda}_i(\boldsymbol{\lambda})$ for a parameter vector $\boldsymbol{\lambda}$. This generalization still fits into the theoretical framework presented in section 3. Nonetheless, for the sake of notational simplicity, we restrict ourselves to assumption (i). As indicated by the index i, the $\mathbf{b}_i$ are different among the groups. All observations have the coefficient $\boldsymbol{\beta}$ in common whereas the value of $\mathbf{b}_i$ depends on the group that the observation belongs to. In other words, for each group there are group-specific deviations $\mathbf{b}_i$ from the overall effects $\boldsymbol{\beta}$. We assume throughout the article that the design matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ are deterministic, that is, fixed design. The case with random design does neither change the methodology (i.e. we can use the same methods presented next) nor alter much the theory, although the latter needs an adaptation of mathematical arguments.

We allow that the number p of fixed-effects regression coefficients may be much larger than the total number of observations, that is, $N_T\ll p$. Furthermore, the number q of random-effects variables might be as large as $q\leq p$, but the dimension q^* of the variance–covariance parameters is assumed to be small ($q^*\ll N_T$). We aim at estimating the fixed regression parameter vector $\boldsymbol{\beta}$, the random effects $\mathbf{b}_i$ and the variance–covariance parameters $\boldsymbol{\theta}$ and σ^2. Therefore,

$\tilde{\boldsymbol{\phi}} := (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \sigma^2)^T$ defines the complete parameter vector with at most length $p + q(q+1)/2 + 1$. From model (5) we deduce that $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are independent and $\mathbf{y}_i \sim \mathcal{N}_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i(\boldsymbol{\theta}, \sigma^2))$ with $\mathbf{V}_i(\boldsymbol{\theta}, \sigma^2) = \mathbf{Z}_i\boldsymbol{\Psi}_{\boldsymbol{\theta}}\mathbf{Z}_i^T + \sigma^2\mathbf{I}_{n_i}$. Denote the stacked vectors $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_N^T)^T$, $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_N^T)^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \ldots, \boldsymbol{\varepsilon}_N^T)^T$ and the stacked matrices $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_N^T)^T$, $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)$ and $\mathbf{V} = \mathrm{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_N)$. Then model (5) can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon} \tag{6}$$

and the negative log-likelihood is given by

$$-\ell(\tilde{\boldsymbol{\phi}}) = -\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{2}\{N_T \log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\}, \tag{7}$$

where $|\mathbf{V}| = \det(\mathbf{V})$.

2.2. ℓ_1-penalized maximum likelihood estimator

Because of the possibly large number of covariates ($N_T \ll p$ setting), we cannot use the classical maximum likelihood or restricted maximum likelihood (REML) 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:

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) := \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{k=1}^{p}|\beta_k|, \tag{8}$$

where λ is a non-negative regularization parameter. Consequently, we estimate the fixed regression coefficient vector $\boldsymbol{\beta}$ and the variance components $\boldsymbol{\theta}$ and σ^2 by

$$\hat{\tilde{\boldsymbol{\phi}}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 > 0, \boldsymbol{\Psi} > 0}{\arg\min} \; Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2). \tag{9}$$

For fixed variance parameters $\boldsymbol{\theta}$ and σ^2, the minimization with respect to $\boldsymbol{\beta}$ is a convex optimization problem. Since we want to make use of this convexity (see section 4), we do not profile the likelihood function, as usually performed in the mixed-effects model framework (Pinheiro & Bates, 2000). However, with respect to the full parameter vector $\tilde{\boldsymbol{\phi}}$, 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 discuss how to address this issue.

2.3. Prediction of the random-effects coefficients

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

$$\begin{aligned}
\tilde{\mathbf{b}}_i &= \underset{\mathbf{b}_i}{\arg\max} f(\mathbf{b}_i \mid \mathbf{y}_1, \ldots, \mathbf{y}_N, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \underset{\mathbf{b}_i}{\arg\max} f(\mathbf{b}_i \mid \mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \\
&= \underset{\mathbf{b}_i}{\arg\max} \frac{f(\mathbf{y}_i \mid \mathbf{b}_i, \boldsymbol{\beta}, \sigma^2) \cdot f(\mathbf{b}_i \mid \boldsymbol{\theta})}{f(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2)} \\
&= \underset{\mathbf{b}_i}{\arg\min} \left\{ \frac{1}{\sigma^2}\|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i\|^2 + \mathbf{b}_i^T\boldsymbol{\Psi}_{\boldsymbol{\theta}}^{-1}\mathbf{b}_i \right\}.
\end{aligned}$$

From this we get $\tilde{\mathbf{b}}_i = [\mathbf{Z}_i^T\mathbf{Z}_i + \sigma^2\boldsymbol{\Psi}_{\boldsymbol{\theta}}^{-1}]^{-1}\mathbf{Z}_i^T\mathbf{r}_i$, where $\mathbf{r}_i = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ is the (marginal) residual vector. Since the true values of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ^2 are unknown, the $\mathbf{b}_i$s are predicted by $\hat{\mathbf{b}}_i = [\mathbf{Z}_i^T\mathbf{Z}_i + \hat{\sigma}^2\boldsymbol{\Psi}_{\hat{\boldsymbol{\theta}}}^{-1}]^{-1}\mathbf{Z}_i^T\hat{\mathbf{r}}_i$ with $\hat{\mathbf{r}}_i = (\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$, using the estimates from (9).

2.4. Selection of the regularization parameter

The estimation requires to choose a regularization parameter λ. We propose to use the Bayesian information criterion (BIC) defined by

$$\mathrm{BIC}_\lambda := -2\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) + \log N_T \cdot \hat{df}_\lambda, \tag{10}$$

where $\hat{df}_\lambda := |\{1 \le k \le p : \hat{\beta}_k \neq 0\}| + \dim(\boldsymbol{\theta})$ is the sum of the number of the non-zero fixed regression coefficients and the number of variance–covariance parameters. The use of $|\{1 \le k \le p : \hat{\beta}_k \neq 0\}|$ 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 non-zero estimated coefficients.

Obviously, there are other tuning parameter selection methods, for example cross-validation and Akaike information-type criteria, among others. Advocating the BIC as selection criterion is based on our empirical experience that it performs best in both simulations and real data examples (see sections 5 and 6).

2.5. Adaptive ℓ_1-penalized maximum likelihood estimator

Because of the bias of the Lasso, Zou (2006) proposed the adaptive Lasso. For some given weights $w_1, \ldots, w_p$, the adaptive ℓ_1-penalized maximum likelihood estimator has the following objective function instead of (8):

$$Q_\lambda^{w_1,\ldots,w_p}(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) := \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{k=1}^{p} w_k|\beta_k|,$$

and hence

$$\hat{\hat{\boldsymbol{\phi}}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 > 0, \boldsymbol{\Psi} > 0}{\arg\min}\ Q_\lambda^{w_1,\ldots,w_p}(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2). \tag{11}$$

The weights $w_1, \ldots, w_p$ may be calculated from an initial estimator $\hat{\boldsymbol{\beta}}_{\mathrm{init}}$ in (9) with $w_k := 1/|\hat{\beta}_{\mathrm{init},k}(\lambda)|$ for $k = 1, \ldots, p$. Unless specified otherwise, we employ these weights.

3. Theoretical results

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 (8) and (9) we see that we are dealing with a non-convex loss function, because of the variance parameters $\boldsymbol{\theta}$ and σ^2, and a convex ℓ_1-penalty. To the best of our knowledge, only Städler *et al.* (2010) consider high-dimensional non-convex ℓ_1-penalized smooth likelihood problems. In this section, we build upon the theory presented in Städler *et al.* (2010) and extend their results to prove an oracle inequality for the adaptive ℓ_1-penalized estimator (11).

We use the following framework and notation. Let $i = 1, \ldots, N$ as before. In the remainder of this article, we assume that $n_i \equiv n > 1$ the same for all i. This assumption eases the mathematical presentation while all theoretical and computational results remain valid for different n_is. Denote by $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^n$ the response variable. Let $\mathbf{X}_i$ be the fixed covariates in

some space $\mathcal{X}^n \subset \mathbb{R}^{n\times p}$ and $\mathbf{Z}_i \subset \mathbf{X}_i$. 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 $\boldsymbol{\phi}^T := (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, 2\log\sigma) = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \varrho) = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{p+q^*+1}$ and denote by $\boldsymbol{\phi}_0$ the true parameter vector. For a constant $0 < K < \infty$, consider the parameter space

$$\boldsymbol{\Phi} = \{\boldsymbol{\phi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) : \sup_{\mathbf{x}\in\mathcal{X}} |\mathbf{x}^T\boldsymbol{\beta}| \le K, \|\boldsymbol{\eta}\|_\infty \le K, \boldsymbol{\Psi} > 0\} \in \mathbb{R}^{p+q^*+1}, \tag{12}$$

where $\|\boldsymbol{\eta}\|_\infty = \max_l |\eta_l|$. We modify the estimators in (9) and (11) by restricting the solution to lie in the parameter space $\boldsymbol{\Phi}$:

$$\hat{\boldsymbol{\phi}} := \arg\min_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \left\{ \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{k=1}^{p}|\beta_k| \right\}, \tag{13}$$

$$\hat{\boldsymbol{\phi}}_{\text{weight}} := \arg\min_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} \left\{ \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\sum_{k=1}^{p} w_k|\beta_k| \right\}. \tag{14}$$

Now, let $f_{\boldsymbol{\phi},\mathbf{X}_i,\mathbf{Z}_i}$ be the Gaussian density for $\mathbf{y}_i$ with respect to the aforegiven parameterization. Since we use the negative log-likelihood as loss function, the excess risk coincides with the Kullback–Leibler distance:

$$\mathcal{E}_{\mathbf{X},\mathbf{Z}}(\boldsymbol{\phi}\,|\,\boldsymbol{\phi}_0) = \int \log\left(\frac{f_{\boldsymbol{\phi}_0,\mathbf{X},\mathbf{Z}}}{f_{\boldsymbol{\phi},\mathbf{X},\mathbf{Z}}}\right) f_{\boldsymbol{\phi}_0,\mathbf{X},\mathbf{Z}}\,\mathrm{d}\mu, \tag{15}$$

where μ denotes the Lebesgue measure, and we define the average excess risk as:

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

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

3.1. Consistency for the ℓ_1-penalized estimator

We require only one condition for consistency. It is a condition on the random-effects design matrices $\mathbf{Z}_i$.

Assumption 1. *The eigenvalues of $\mathbf{Z}_i^T\mathbf{Z}_i$, denoted by $(v_j^{(i)})_{j=1}^q$ for $i = 1,\ldots,N$, are bounded: $v_j^{(i)} \le K < \infty$ for all i and j, with K from* (12).

Now we consider a triangular scheme of observations from (5):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_N + \mathbf{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1,\ldots,N, \tag{16}$$

where the parameters $\boldsymbol{\beta}_N$ and $\boldsymbol{\eta}_N$ are allowed to depend on N. We study consistency as $N\to\infty$ but the group size n is fixed (in general, $(n_i)_{i\ge1}$ are fixed). Moreover, let us use the notation $a\vee b := \max\{a,b\}$.

Theorem 1 (*consistency*)**.** *Consider model* (16) *and the estimator* (13). *Under assumption* 1 *and assuming*

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

for some $C > 0$, any global minimizer $\hat{\boldsymbol{\phi}}$ as in (13) *satisfies $\bar{\mathcal{E}}(\hat{\boldsymbol{\phi}}\,|\,\boldsymbol{\phi}_0) = o_P(1)$ as $N\to\infty$.*

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

3.2. Oracle inequality for the adaptive ℓ_1-penalized estimator

We now present an oracle optimality result in non-asymptotic form for the adaptive estimator (and thereby covering also the non-adaptive case). Preliminarily, we introduce some notation and two further assumptions.

Assumption 2.

(a) *Let* $(\omega_j^{(i)})_{j=1}^n$ *be the eigenvalues of* $\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^T$ *for* $i=1,\ldots,N$. *At least two eigenvalues are different, that is, for all* $i\ \exists\, j_1\neq j_2\in\{1,\ldots,n\}$ *such that* $\omega_{j_1}^{(i)}\neq\omega_{j_2}^{(i)}$.

(b) *For* $i=1,\ldots,N$, *the matrices* $\boldsymbol{\Omega}_i$ *defined by*

$$(\boldsymbol{\Omega}_i)_{r,s}=\operatorname{tr}\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\phi_{p+r}}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\phi_{p+s}}\right)\quad r,s=1,\ldots,q^*+1$$

are strictly positive definite.

Remark. In the special case $\boldsymbol{\Psi}=\theta^2\mathbf{I}$, assumption 2(b) automatically holds.

Let $S(\boldsymbol{\beta})=\{1\leq k\leq p:\beta_k\neq 0\}$ be the active set of $\boldsymbol{\beta}$, that is, the set of non-zero coefficients, and $\boldsymbol{\beta}_{\mathcal{K}}=\{\beta_k:k\in\mathcal{K}\}$ for $\mathcal{K}\subset\{1,\ldots,p\}$. We denote by $S_0=S(\boldsymbol{\beta}_0)$ the true active set and by $s_0=|S_0|$ its cardinality. Write $\mathbf{X}_i^T=(\mathbf{x}_1^i,\ldots,\mathbf{x}_n^i)$ and define

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

Assumption 3 (*restricted eigenvalue condition*)**.** *There exists a constant* $\kappa\geq 1$, *such that for all* $\boldsymbol{\beta}\in\mathbb{R}^p$ *satisfying* $\|\boldsymbol{\beta}_{S_0^c}\|_1\leq 6\|\boldsymbol{\beta}_{S_0}\|_1$ *it holds that* $\|\boldsymbol{\beta}_{S_0}\|_2^2\leq\kappa^2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{N,n}\boldsymbol{\beta}$.

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

$$\lambda_0=M_N\log N\sqrt{\frac{\log(p\vee N)}{N}},\tag{17}$$

where M_N is of order $\log N$ and an exact definition is given in the proof of theorem 1. For any $T\geq a_1$, let $\mathcal{J}$ be a set defined by the underlying empirical process (see (A.6) in the Supporting Information). It is shown in the proof of theorem 1 that the set $\mathcal{J}$ has large probability,

$$\mathbb{P}[\mathcal{J}]\geq 1-a_2\exp\left[-\frac{T^2\log^2 N\log(p\vee N)}{a_3^2}\right]-\frac{\rho}{\log N}\frac{1}{N^{1-2\varepsilon}}$$

for N sufficiently large and some constants $a_1,a_2,a_3,\varepsilon,\rho>0$; see lemmas 2 and 3 in appendix A (in the Supporting Information).

At this point, we could conclude an oracle result in the way of Städler *et al.* (2010). However, we extend that result and present an oracle inequality involving $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_1$ instead of $\|\hat{\boldsymbol{\beta}}_{S_0^c}\|_1$ for the ℓ_1-penalized as well as the adaptive ℓ_1-penalized estimator.

Theorem 2 (*oracle result*)**.** *Consider the weighted* ℓ_1*-penalized estimator* (14). *Suppose that for some* $\delta>0$,

$$w_k \begin{cases} \leq 1/\delta, & k \in S_0, \\ \geq 1/\delta, & k \notin S_0. \end{cases}$$

Under assumptions 1, 2 *and* 3, *and for* $\lambda \geq 2T\delta\lambda_0$, *we have on the set* $\mathcal{J}$ *defined in* (A.6),

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\phi}}_{\text{weight}} \,|\, \boldsymbol{\phi}_0) + 2(\lambda/\delta - T\lambda_0)\|\hat{\boldsymbol{\beta}}_{\text{weight}} - \boldsymbol{\beta}_0\|_1 \leq 9(\lambda/\delta + T\lambda_0)^2 c_0^2 \kappa^2 s_0.$$

The proof is given in appendix B. Application of theorem 2 with all weights equal to one ($\delta = 1$) gives an oracle result for the ℓ_1-penalized estimator, which we will use as initial values for the adaptive Lasso procedure.

Corollary 1. *Let*

$$\hat{\boldsymbol{\phi}}_{\text{init}} := (\hat{\boldsymbol{\beta}}_{\text{init}}, \hat{\boldsymbol{\theta}}_{\text{init}}, \hat{\varrho}_{\text{init}}) := \arg\min_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} Q^{1,\ldots,1}_{\lambda_{\text{init}}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \varrho),$$

be the initial estimator in (14) (*i.e. the estimator with all the weights equal to* 1). *Under assumptions* 1, 2 *and* 3, *and for* $\lambda_{\text{init}} \geq 2T\lambda_0$, *we have on* $\mathcal{J}$,

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\phi}}_{\text{init}} \,|\, \boldsymbol{\phi}_0) + 2(\lambda_{\text{init}} - T\lambda_0)\|\hat{\boldsymbol{\beta}}_{\text{init}} - \boldsymbol{\beta}_0\|_1 \leq 9(\lambda_{\text{init}} + T\lambda_0)^2 c_0^2 \kappa^2 s_0. \tag{18}$$

It is clear that the ℓ_1-estimation error bound implies a bound for the ℓ_∞ estimation error as well. When the underlying true coefficients $\beta_{0,k}$, $k \in S_0$ are sufficiently much larger in absolute value than the ℓ_∞-estimation error bound, one can perfectly distinguish between the active and non-active sets. This argument is applied in the next corollary to the adaptive Lasso with estimated weights.

Corollary 2. *Let*

$$\hat{\boldsymbol{\phi}}_{\text{adap}} := (\hat{\boldsymbol{\beta}}_{\text{adap}}, \hat{\boldsymbol{\theta}}_{\text{adap}}, \hat{\varrho}_{\text{adap}}) := \arg\min_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} Q^{w_1,\ldots,w_p}_{\lambda_{\text{adap}}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \varrho),$$

be the adaptive estimator with weights $w_k = 1/|\hat{\beta}_{\text{init},k}|$, $k = 1, \ldots, p$ *as in* (13). *Assume that for all* $k \in S_0$,

$$|\beta_{0,k}| \geq 2\delta_{\text{init}}, \tag{19}$$

where

$$\delta_{\text{init}} := \frac{9(\lambda_{\text{init}} + T\lambda_0)^2 c_0^2 \kappa^2 s_0}{2(\lambda_{\text{init}} - T\lambda_0)} \geq \|\hat{\boldsymbol{\beta}}_{\text{init}} - \boldsymbol{\beta}_0\|_1$$

is a bound of the ℓ_1*-estimation error of the initial* $\hat{\boldsymbol{\beta}}_{\text{init}}$. *Suppose moreover that assumptions* 1, 2 *and* 3 *are met. Then, for* $\lambda_{\text{adap}} \geq 2T\delta_{\text{init}}\lambda_0$, *and on the set* $\mathcal{J}$,

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\phi}}_{\text{adap}} \,|\, \boldsymbol{\phi}_0) + 2(\lambda_{\text{adap}}/\delta_{\text{init}} - T\lambda_0)\|\hat{\boldsymbol{\beta}}_{\text{adap}} - \boldsymbol{\beta}_0\|_1 \leq 9(\lambda_{\text{adap}}/\delta_{\text{init}} + T\lambda_0)^2 c_0^2 \kappa^2 s_0. \tag{20}$$

We call condition (19) a 'betamin' condition. It is clearly very restrictive, but allows for an easy derivation of the oracle result. The 'betamin' condition can indeed be substantially refined. In van de Geer *et al.* (2010), one can find similar oracle results, and in addition variable selection results, for the adaptive Lasso in the linear model, without 'betamin' conditions. These results require introducing various versions of restricted eigenvalues and sparse eigenvalues, and can be generalized to the current setting. Since a full presentation is rather involved, we have confined ourselves to the simplest case.

Recall that in (17), we choose λ_0 of order $\log^2 N\sqrt{\log(p \vee N)/N}$. When we also choose λ_{init} of this order, we find, modulo the restricted eigenvalue κ and the constants T and c_0, that the right-hand side of the oracle result (18) for the initial estimator is of order

$$\log^4 N \frac{\log(p \vee N)}{N} s_0,$$

and that

$$\delta_{\text{init}} \asymp \log^2 N \sqrt{\frac{\log(p \vee N)}{N}} s_0.$$

The tuning parameter for the adaptive Lasso can then be taken of order

$$\lambda_{\text{adap}} \asymp \log^4 N \frac{\log(p \vee N)}{N} s_0.$$

The right-hand side (20) of the oracle result for the adaptive estimator is then of the same order as that for the initial estimator.

Assuming 'betamin' conditions, the results in corollaries 1 and 2 imply the variable screening property motivated already in (4).

Corollary 3.

(1) *For the* ℓ_1*-penalized* (*initial*) *estimator* (13), *assume*

$$\min_k |\beta_{0,k}| > \delta_{\text{init}} = \frac{9(\lambda_{\text{init}} + T\lambda_0)^2 c_0^2 \kappa^2 s_0}{2(\lambda_{\text{init}} - T\lambda_0)}.$$

Then, under the assumptions of corollary 1, *on the set* $\mathcal{J}$,

$$S_0 \subset \hat{S}_{\text{init}} = \{1 \le k \le p : \hat{\beta}_{\text{init},k} \neq 0\}.$$

(2) *For the adaptive* ℓ_1*-penalized estimator in corollary* 2, *assume*

$$\min_k |\beta_{0,k}| > \frac{9(\lambda_{\text{adap}}/\delta_{\text{init}} + T\lambda_0)^2 c_0^2 \kappa^2 s_0}{2(\lambda_{\text{adap}}/\delta_{\text{adap}} - T\lambda_0)}.$$

Then, under the assumptions of corollary 2, *on the set* $\mathcal{J}$,

$$S_0 \subset \hat{S}_{\text{adap}} = \{1 \le k \le p : \hat{\beta}_{\text{adap},k} \neq 0\}.$$

The proof of corollary 3 *is given in appendix* B.

4. Computational algorithm

The algorithm for the estimators in (9) and (11) is based on the Block Coordinate Gradient Descent (BCGD) method from Tseng & Yun (2009). The main ideas of our BCGD algorithm are that we cycle through the coordinates and minimize the objective function $Q_\lambda(\cdot)$ with respect to only one coordinate while keeping the other parameters fixed (i.e. a Gauss–Seidel algorithm). In each such step, we approximate $Q_\lambda(\cdot)$ 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 group Lasso as well as in Wu & Lange (2008) and Friedman *et al.* (2010) for the ordinary Lasso. We remark that Meier *et al.* (2008) have a block structure because of 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–covariance parameters.

Let us first introduce the notation and give an overview of the algorithm before proving that our optimization problem achieves numerical convergence. All the details as well as

some computational aspects are deferred to appendix C in the Supporting Information on the journal website.

Let $\boldsymbol{\phi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{p+q^*+1}$ be the parameterization introduced in the previous section. Define the functions

$$P(\boldsymbol{\phi}) := \sum_{k=1}^{p} |\beta_k|, \quad g(\boldsymbol{\phi}) := \frac{1}{2}\log|\mathbf{V}(\boldsymbol{\eta})| + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\boldsymbol{\eta})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}).$$

Now (9) can be written as $\hat{\boldsymbol{\phi}}_\lambda = \arg\min_{\boldsymbol{\phi}} Q_\lambda(\boldsymbol{\phi}) := g(\boldsymbol{\phi}) + \lambda P(\boldsymbol{\phi})$. Letting $\mathbf{e}_j$ be the jth unit vector, the algorithm can be summarized in the following way:

Algorithm 1 (coordinate gradient descent).

(0) *Let* $\boldsymbol{\phi}^0 \in \mathbb{R}^{p+q^*+1}$ *be an initial value.*
For $\ell = 0, 1, 2, \ldots,$ *let* $\mathcal{S}^\ell$ *be the index cycling through the coordinates* $\{1\}, \{2\}, \ldots, \{p+q^*\}, \{p+q^*+1\}$

(1) *Approximate the second derivative* $\partial^2/\partial(\phi_{\mathcal{S}^\ell})^2 Q_\lambda(\boldsymbol{\phi}^\ell)$ *by* $h^\ell > 0$.

(2) *Calculate the descent direction*

$$d^\ell := \arg\min_{d\in\mathbb{R}} \left\{ g(\boldsymbol{\phi}^\ell) + \frac{\partial}{\partial\phi_{\mathcal{S}^\ell}} g(\boldsymbol{\phi}^\ell)d + \frac{1}{2}d^2h^\ell + \lambda P(\boldsymbol{\phi}^\ell + d\mathbf{e}_{\mathcal{S}^\ell}) \right\}.$$

(3) *Choose a stepsize* $\alpha^\ell > 0$ and set $\boldsymbol{\phi}^{\ell+1} = \boldsymbol{\phi}^\ell + \alpha^\ell d^\ell \mathbf{e}_{\mathcal{S}^\ell}$ *such that there is a decrease in the objective function;*

until convergence.

The details of (0)–(3) and further computational issues are given in appendix C of the online Supporting Information. An implementation of the algorithm can be found in the R package lmmlasso, which is available from R-Forge (http://lmmlasso.R-forge.R-project.org).

The convergence properties of the CGD algorithm are described in the following theorem.

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

The proof is given in appendix C.

In general, because of the non-convexity of the optimization problem, the CGD algorithm may not achieve the global optimum.

5. Simulation study

In this section, we assess the empirical performance of the ℓ_1-penalized maximum likelihood estimators (9) and (11) in different kinds of simulation examples. We study several performance measures and compare the proposed method with Lasso and linear mixed-effects methods, if possible. After some introductory remarks, we focus on high-dimensional examples. The simulation study for the low-dimensional setting is provided in the online Supporting Information. The application of the new procedure on a real data set is illustrated in the next section.

Hereafter, we denote by *lmmLasso* the ℓ_1-penalized maximum likelihood estimator (9), by *lmmadLasso* the adaptive ℓ_1-penalized maximum likelihood estimator (11) and by *lme* the classical linear mixed-effects model provided by the R package nlme (Pinheiro & Bates, 2000). Furthermore, let *Lasso* denote the standard Lasso (Efron *et al.*, 2004) and *adLasso*

the adaptive Lasso (Zou, 2006) where the regularization parameter is chosen by minimizing the BIC.

As an overview, let us summarize the most important conclusions from the simulation studies:

(a) The variability of the estimated fixed-effects parameters $\hat{\beta}_k$ is much smaller if there is no corresponding random effect $(\hat{b}_i)_k$ for $i=1,\ldots,N$, for all *lme*, *lmmLasso* and *lmmadLasso*.

(b) In the high-dimensional framework, the following aspects appear (and are virtually not observable in the low-dimensional setting):
   1. Penalizing fixed-effects covariates which also incorporate a random effect causes bias problems. To be more specific, let us assume that the penalized kth covariate has a fixed and a random-effects coefficient, that is, β_k and $(b_i)_k$, respectively. If the regularization parameter λ is large and β_k subject to penalization, then $\hat{\beta}_k$ is shrunken towards zero. Thereby, the estimate of the corresponding variance parameter gets large and $(\hat{b}_i)_k$ has a bias related to the amount of shrinkage in $\hat{\beta}_k$. As a consequence, covariates with fixed and random effect should no be subject to penalization.
   2. An adaptive procedure (11) with appropriate weights may reduce this adverse effect, but it does not overcome the aforementioned problem completely. The work of Bondell *et al.* (2010) covers only the low-dimensional case and the authors do not present any parameter estimates in the simulation study.
   3. The results are sensitive to the starting value. Nonetheless, choosing the starting value as described in appendix C of the online Supporting Information performs well in all simulation examples.

(c) There is a remarkable reduction of the estimated error variance $\hat{\sigma}^2$ when incorporating the random-effects structure in *lmmLasso*, *lmmadLasso* and *lme* compared with *Lasso* and *adLasso*.

(d) The variability of the *Lasso* and *adLasso* coefficient estimators are larger than the corresponding variability of the mixed-effects model approaches.

(e) If we focus on the identification of random-effects covariates, we suggest using a diagonal structure for $\mathbf{\Psi}$ and then eliminating those random-effects covariates with a small variance. An elaborate discussion of the selection of the random-effects structure is beyond the scope of this article. In section 6, we suggest a strategy how to remedy this problem.

In all subsequent simulation schemes, we restrict ourselves to the case where all groups have the same number of observations, that is, we set $n_i \equiv n$ for $i=1,\ldots,N$. Let the first column of $\mathbf{X}_i$ be the (non-penalized) intercept. We assign $\mathbf{Z}_i \subset \mathbf{X}_i$ such that the columns of $\mathbf{Z}_i$ correspond to the first q columns of $\mathbf{X}_i$. This means that the first q variables have both a fixed-effects coefficient β_k and a random-effects coefficient $(b_i)_k$ for $i=1,\ldots,N$ and $k=1,\ldots,q$. The covariates are generated from a multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$ with the pairwise correlation $\mathbf{\Sigma}_{kk'}=\rho^{|k-k'|}$ and $\rho=0.2$. Denote by $\boldsymbol{\beta}_0$ the true fixed effects and by $s_0 := \#\{1 \le k \le p : \beta_{0,k} \neq 0\}$ the true number of non-zero coefficients. Unless otherwise stated, we set $\mathbf{\Psi}=\theta^2\mathbf{I}$. In all subsequent tables, a non-penalized fixed-effects coefficient is marked by an asterisk $(*)$.

5.1. High-dimensional setting

We study four examples in the high-dimensional setting ($\beta_{0,1}=1$ is the unpenalized intercept).

H_1: $N=25$, $n=6$, $N_T=150$, $p=300$, $q=2$, $\sigma^2=0.25$, $\theta^2=0.56$ and $s_0=5$ with $\boldsymbol{\beta}_0=(1,2,4,3,3,0,\ldots,0)^T$.

H_2: $N=30$, $n=6$, $N_T=180$, $p=500$, $q=1$, $\sigma^2=0.25$, $\theta^2=0.56$ and $s_0=5$ with $\boldsymbol{\beta}_0=(1,2,4,3,3,0,\ldots,0)^T$.

H_3: $N=30$, $n=6$, $N_T=180$, $p=1000$, $q=3$, $\sigma^2=0.25$, $\theta^2=0.56$ and $s_0=5$ with $\boldsymbol{\beta}_0=(1,2,4,3,3,0,\ldots,0)^T$.

H_4: $N=25$, $n=6$, $N_T=150$, $p=300$, $\sigma^2=0.25$,

$$\boldsymbol{\Psi}=\begin{pmatrix}3&0&0\\0&3&0\\0&0&2\end{pmatrix}$$

and $s_0=5$ with $\boldsymbol{\beta}_0=(1,2,4,3,3,0,\ldots,0)^T$. In contrast to the previous examples, we fit a wrong model assuming that $\boldsymbol{\Psi}$ is diagonal with dimension 4.

The results in the form of means and standard deviations (in parentheses) over 100 simulation runs are depicted in Tables 2–4. Therein, $|S(\hat{\boldsymbol{\beta}})|$ denotes the cardinality of the estimated active set and TP is the number of true positives.

Let us sum up the simulation results for the models H_1–H_4. As in the low-dimensional setting (see appendix D), the estimated active set is sparse and all methods include the true non-zero coefficients.

Table 2 reveals that *lmmLasso* and *lmmadLasso* reduce the error variance remarkably in comparison with *Lasso* and *adLasso*. Nevertheless, *lmmLasso* overestimates the true value of σ^2 whereas *lmmadLasso* underestimates σ^2. We observe, in particular for H_4, that a maximum

Table 2. *Simulation results for* H_1, H_2 *and* H_3 (*an asterisk indicates that the fixed-effects coefficient is not subject to penalization*)

	Method	$\lvert S(\hat{\boldsymbol{\beta}})\rvert$	TP	$\hat{\sigma}^2$	$\hat{\theta}^2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Model	True	5	5	0.25	0.56	1	2	4	3	3
H_1	lmmLasso	6.70	5	0.29	0.52	1.01*	2.05*	3.86	2.90	2.88
		(2.14)	(0)	(0.05)	(0.12)	(0.16)	(0.16)	(0.06)	(0.06)	(0.06)
	lmmadLasso	6.59	5	0.22	0.52	1.01*	2.03*	3.98	2.99	3.00
		(2.02)	(0)	(0.04)	(0.12)	(0.16)	(0.16)	(0.06)	(0.05)	(0.05)
	Lasso	6.29	5	1.36	–	1.01*	2.07*	3.76	2.84	2.79
		(1.46)	(0)	(0.27)	–	(0.17)	(0.19)	(0.10)	(0.11)	(0.10)
	adLasso	6.29	5	1.16	–	1.01*	2.02*	3.98	3.00	2.99
		(1.46)	(0)	(0.24)	–	(0.17)	(0.18)	(0.10)	(0.11)	(0.10)
H_2	lmmLasso	6.65	5	0.28	0.56	1.00*	1.90	3.91	2.92	2.89
		(1.71)	(0)	(0.04)	(0.17)	(0.15)	(0.04)	(0.05)	(0.04)	(0.05)
	lmmadLasso	6.53	5	0.22	0.55	1.00*	2.00	3.99	3.00	2.99
		(1.64)	(0)	(0.03)	(0.17)	(0.15)	(0.04)	(0.04)	(0.04)	(0.04)
	Lasso	6.84	5	0.87	–	1.00*	1.84	3.88	2.88	2.83
		(2.02)	(0)	(0.19)	–	(0.15)	(0.08)	(0.07)	(0.09)	(0.08)
	adLasso	6.84	5	0.72	–	1.00*	2.00	4.00	3.00	2.98
		(2.02)	(0)	(0.17)	–	(0.15)	(0.07)	(0.07)	(0.08)	(0.08)
H_3	lmmLasso	6.17	5	0.29	0.52	1.02*	2.00*	4.04*	2.84	2.84
		(1.74)	(0)	(0.05)	(0.10)	(0.15)	(0.15)	(0.15)	(0.07)	(0.06)
	lmmadLasso	6.12	5	0.23	0.53	1.02*	2.00*	4.00*	2.99	2.99
		(1.70)	(0)	(0.04)	(0.10)	(0.15)	(0.15)	(0.15)	(0.07)	(0.06)
	Lasso	5.93	5	1.94	–	1.03*	2.02*	4.06*	2.70	2.70
		(1.48)	(0)	(0.36)	–	(0.17)	(0.18)	(0.19)	(0.11)	(0.13)
	adLasso	5.93	5	1.69	–	1.03*	2.02*	3.99*	2.98	2.97
		(1.48)	(0)	(0.32)	–	(0.16)	(0.17)	(0.18)	(0.12)	(0.12)

TP, true positives; Lasso, least absolute shrinkage and selection operator.

Table 3. *Simulation results for H_4 (an asterisk indicates that the fixed-effects coefficient is not subject to penalization)*

Method	$\lvert S(\hat{\boldsymbol{\beta}})\rvert$	TP	$\hat{\sigma}^2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
True	5	5	0.25	1	2	4	3	3
lmmLasso	5.56	5	0.26	0.95*	1.99*	3.97*	3.04*	2.82
	(0.97)	(0)	(0.05)	(0.31)	(0.38)	(0.31)	(0.07)	(0.07)
lmmadLasso	5.56	5	0.22	0.95*	1.99*	3.97*	3.00*	3.00
	(0.97)	(0)	(0.04)	(0.31)	(0.38)	(0.30)	(0.07)	(0.06)
Lasso	6.84	5	7.85	0.94*	2.01*	3.99*	3.11*	2.36
	(12.18)	(0)	(1.81)	(0.38)	(0.47)	(0.38)	(0.23)	(0.28)
adLasso	6.79	5	7.25	0.95*	2.02*	4*	2.98*	3.01
	(11.68)	(0)	(1.76)	(0.37)	(0.47)	(0.38)	(0.22)	(0.29)

TP, true positives; Lasso, least absolute shrinkage and selection operator.

Table 4. *Mean covariance estimates for H_4*

Method	Ψ_{11}	Ψ_{22}	Ψ_{33}	Ψ_{44}
True	3	3	2	0
lmmLasso	2.82	2.94	1.85	0.01
	(0.80)	(0.88)	(0.62)	(0.02)
lmmadLasso	2.81	2.94	1.84	0.01
	(0.79)	(0.88)	(0.62)	(0.02)

Lasso, least absolute shrinkage and selection operator.

likelihood approach (in contrast to a restricted maximum likelihood approach) gives biased variance–covariance estimators. It is possible to implement an REML-type approach (Ni *et al.*, 2010) in the high-dimensional setting to reduce the bias in the variance parameters. However, we have observed that (i) the number of (Gauss–Seidel) cycles increases and (ii) the algorithm may fail to converge. In all models, we do not penalize the covariates with both a fixed and random effect. Without doing this (not shown here), the fixed effects would be set to zero whereas the estimated between-subject variability $\hat{\theta}^2$ would increase. As a consequence, the predicted random effects are too large and are not centered at zero, but around the true fixed effect. Hence this would result in a model which does not fulfill the assumptions in (5) anymore.

Tables 2 and 3 reveal that the variability of the fixed effects with no corresponding random effect is approximately half of the non-penalized coefficients. This difference of estimation variability is also observed in the classical linear mixed-effects framework (see *lme* in Tables 7 and 8 in appendix D). Besides, *lmmLasso* has a bias towards zero, which is notably smaller than that from the *Lasso*. As expected, this bias can be reduced by *lmmadLasso*. Concerning H_4, it is worth pointing out that although not knowing the true covariance structure, we may use a diagonal structure for $\boldsymbol{\Psi}$ and then drop the variances which are close to zero. A suggestion how to use this idea in a real data set is presented in the next section.

5.2. Within-group prediction performance

We now turn to consider the performance of the proposed methodology concerning within-group prediction. We compare the predictive performance between six different Lasso procedures. In doing so, denote by *lmmLasso*, *lmmadLasso*, *Lasso* and *adLasso* the procedures from the previous subsection. In addition, let *cv-Lasso* be a cross-validated Lasso

Table 5. *Mean squared prediction error for three simulation examples*

Model	θ^2	lmmLasso	lmmad Lasso	Lasso	adLasso	cv-Lasso	cv-ad Lasso
$P_1(p=10)$	0	1.01	1.02	1.00	1.01	1.05	1.01
	0.25	1.33	1.29	1.76	1.84	1.81	1.84
	1	1.66	1.55	3.74	3.74	3.88	3.77
	2	1.67	1.80	5.92	6.25	5.94	6.25
$P_2(p=100)$	0	1.12	1.02	1.26	1.09	1.20	1.14
	0.25	1.51	1.38	1.75	1.75	2.06	1.75
	1	1.94	1.86	4.35	4.53	4.61	4.23
	2	2.49	1.95	7.04	7.02	7.09	6.98
$P_3(p=500)$	0	1.22	1.07	1.18	1.26	1.24	1.58
	0.25	1.83	1.58	2.63	2.67	2.98	3.58
	1	2.00	1.85	4.35	3.78	4.14	4.85
	2	2.54	2.04	10.30	8.26	9.47	11.28

Lasso, least absolute shrinkage and selection operator.

and *cv-adLasso* a cross-validated adaptive Lasso whose 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 $\boldsymbol{\beta}_0=(1, 1.5, 1.2, 1, 2, 0, \ldots, 0)^T$, $\sigma^2=1$ and $\rho=0.2$. We only alter the number of fixed covariates p and the variance component θ^2. For measuring the quality of prediction, we generate a test set with 50 observations per group and calculate the mean squared prediction error. The three models considered are:

$$P_1 : p=10, \quad P_2 : p=100, \quad P_3 : p=500.$$

The results are shown in Table 5. We see that the methods differ slightly for $\theta^2=0$ which corresponds to no grouping structure. As θ^2 increases, the mean squared prediction error increases less for the *lmmLasso* and the *lmmadLasso* than for the other 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 (5).

6. Application: riboflavin data

6.1. Data description

We illustrate the proposed procedure on a real data set which is provided by DSM (Switzerland). The goal is to genetically modify the bacterium *Bacillus subtilis* to increase its production rate for riboflavin (vitamin B2). A first step to achieve this purpose is to identify genes which are most relevant for the production rate. A linear mixed-effects model is appropriate because several observations at different time points are available (i.e. longitudinal data). The response variable is the logarithm of the riboflavin production rate and there are $p=4088$ covariates (genes) measuring the gene expression levels. We have $N=28$ groups with $n_i \in \{2, \ldots, 6\}$ and $N_T=111$ observations. We standardize all covariates to have mean zero and variance one.

6.2. Model selection strategy

Preliminarily, we address the issue of determining those covariates which have both a fixed and a random-effects coefficient. In other words, we specify the matrices $\mathbf{Z}_i \subset \mathbf{X}_i$. Since we

have to deal with high-dimensional, low sample size data, the various tools proposed in Pinheiro & Bates (2000) for determining $\mathbf{Z}_i$ can hardly be applied. Instead, we suggest the following strategy:

(i) Calculate an ordinary Lasso solution $\hat{\boldsymbol{\beta}}^{\text{Lasso}}$ (with cross-validation) and define the active set $\hat{S}_{\text{init}} := \{1 \le k \le p : \hat{\beta}_k^{\text{Lasso}} \ne 0\}$.

(ii) For each $l \in \hat{S}_{\text{init}}$, fit a model in which only the lth covariate has a random-effects coefficient. Denote the corresponding variance estimate by $\hat{\theta}_l^2$.

(iii) Let $\hat{\theta}_{[1]}^2 \ge \hat{\theta}_{[2]}^2 \ge \cdots \ge \hat{\theta}_{[|S_{\text{init}}|]}^2$ be the ordered estimated variances from (2). Then for $\kappa > 0$ define the set $\mathcal{R}_\kappa := \{l \in S_{\text{init}} : \hat{\theta}_l^2 > \kappa\} \cap \{l \in \hat{S}_{\text{init}} : \text{BIC}_{\hat{\theta}_l^2} \le \text{BIC}_0\}$, where BIC_0 is the BIC of the Lasso solution in (1).

(iv) Fit a model with $\mathbf{Z}_i = \mathbf{X}_i^{\mathcal{R}_\kappa}$ (where $\mathbf{X}_i^{\mathcal{R}_\kappa}$ consists of the variables included in $\mathcal{R}_\kappa$) and $\boldsymbol{\Psi}$ being diagonal and keep the non-zero elements of $\hat{\boldsymbol{\Psi}}$.

By doing so (and setting $\kappa = 0.05$), it seems reasonable to fit a model wherein two covariates have an additional random effect. Denoting these variables as k_1 and k_2, the model can be written as:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_{ik_1} z_{ijk_1} + b_{ik_2} z_{ijk_2} + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, n_i \tag{21}$$

with $b_{ik_1} \sim \mathcal{N}(0, \theta_{k_1}^2)$, $b_{ik_2} \sim \mathcal{N}(0, \theta_{k_2}^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

6.3. Results

We compare the results of *lmmLasso* and *lmmadLasso* with *Lasso* and *adLasso*. The variance component estimates, the cardinality of the active set and the rank R of five fixed-effects coefficients are shown in Table 6. The ranking is determined by ordering the absolute values of the fixed-effects coefficients.

From Table 6, we see that the error variance of the *Lasso* may be considerably reduced by the *lmmLasso*. Although the variance $\hat{\theta}_{k_2}^2$ is small, the BIC of this model is smaller than that of the model including only k_1 as random covariate. It is noteworthy that 53 per cent of the total variability in the data set is because of the between-group variability. This strongly indicates that there is indeed some variability between the groups. As might have been expected from the simulation results, the active set of *lmmLasso* and *lmmadLasso* is smaller than the active set from *Lasso* and *adLasso*. The ranking indicates that there is one dominating covariate whereas the other coefficients differ only slightly between the four procedures (not shown).

Table 6. *Results for lmmLasso, lmmadLasso, Lasso and adLasso of the riboflavin data set*

	lmmLasso	lmmadLasso	Lasso	adLasso
$\hat{\sigma}^2$	0.18	0.15	0.30	0.20
$\hat{\theta}_{k_1}^2$	0.17	0.08	–	–
$\hat{\theta}_{k_2}^2$	0.03	0.03	–	–
$\lvert S(\hat{\boldsymbol{\beta}}) \rvert$	18	14	21	20
$R_{\hat{\beta}_1}$	1	1	1	1
$R_{\hat{\beta}_2}$	2	2	4	6
$R_{\hat{\beta}_3}$	3	3	3	5
$R_{\hat{\beta}_4}$	4	13	–	–
$R_{\hat{\beta}_5}$	5	6	6	7

Lasso, least absolute shrinkage and selection operator.

7. 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 for the adaptive ℓ_1-penalized estimator. 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 error variance can be remarkably reduced when incorporating the knowledge about the cluster structure among observations.

Acknowledgements

The research is supported by the Swiss National Science Foundation (grant no. 20PA21-120043/1). The authors thank the members of the DFG-SNF Forschergruppe 916 for many stimulating discussions.

Supporting Information

Additional Supporting Information can be found in the online version of this article:

Appendix A. Proof of theorem 1 from section 3.
Appendix B. Proof of theorem 2 from section 3.
Appendix C. Computational details of algorithm 1 in section 4.
Appendix D. Simulations for the low-dimensional setting in section 5.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Bickel, P., Ritov, Y. & Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.* **37**, 1705–1732.

Bondell, H. D., Krishna, A. & Gosh, S. K. (2010). Joint variable selection of fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069–1077.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data: methods theory and applications*. Springer-Verlag, New York, Dordrecht, Heidelberg, London.

Bunea, F., Tsybakov, A. & Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. Electron. *J. Statist.* **1**, 169–194.

Candes, E. & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.

Demidenko, E. (2004). *Mixed models, theory and applications*. Wiley, Hoboken, New Jersey.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–499.

Friedman, J., Hastie, T. & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softw.* **33**, 1–22.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36**, 614–645.

van de Geer, S. & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Stat.* **3**, 1360–1392.

van de Geer, S., Bühlmann, P. & Zhou, S. (2010). The adaptive and the thresholded lasso for potentially misspecified models. Preprint arXiv:1001.5176v3.

Greenshtein, E. & Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**, 971–988.
Huang, J., Ma, S. & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statist. Sinica* **18**, 1603–1618.
Ibrahim, J. G., Zhu, H., Garcia, R. I. & Guo, R. (2010). Fixed and random effects selection in mixed effects models. *Biometrics*, no. doi: 10.1111/j.1541-0420.2010.01463.x.
Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **83**, 1014–1022.
Meier, L., van de Geer, S. & Bühlmann, P. (2008). The group lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **70**, 53–71.
Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436–1462.
Meinshausen, N. & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, **37**, 246–270.
Meinshausen, N. & Bühlmann, P. (2010). Stability selection (with discussion). *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **72**, 417–473.
Meinshausen, N., Meier, L. & Bühlmann, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104**, 1671–1681.
Ni, X., Zhang, D. & Zhang, H. H. (2010). Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics* **66**, 79–88.
Osborne, M. R., Presnell, B. & Turlach, B. (2000). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20**, 389–403.
Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. Springer-Verlag, New York.
Städler, N., Bühlmann, P. & van de Geer, S. (2010). l_1-penalization for mixture regression models (with discussion). *TEST* **19**, 209–285.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **58**, 267–288.
Tseng, P. & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.* **117**, 387–423.
Verbeke, G. & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer-Verlag, New York.
Wasserman, L. & Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 2178–2201.
Wu, T. & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2**, 224–244.
Zhang, C.-H. & Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567–1594.
Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563.
Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.
Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173–2192.

Received October 2010, in final form February 2011

Jürg Schelldorfer, Seminar für Statistik, Department of Mathematics, ETH Zurich, CH-8092 Zurich, Switzerland.
E-mail: schelldorfer@stat.math.ethz.ch