Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 374–393
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Relaxed Lasso

Nicolai Meinshausen

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

Available online 17 December 2006

Abstract

The Lasso is an attractive regularisation method for high-dimensional regression. It combines variable selection with an efficient computational procedure. However, the rate of convergence of the Lasso is slow for some sparse high-dimensional data, where the number of predictor variables is growing fast with the number of observations. Moreover, many noise variables are selected if the estimator is chosen by cross-validation. It is shown that the contradicting demands of an efficient computational procedure and fast convergence rates of the ℓ_2-loss can be overcome by a two-stage procedure, termed the relaxed Lasso. For orthogonal designs, the relaxed Lasso provides a continuum of solutions that include both soft- and hard-thresholding of estimators. The relaxed Lasso solutions include all regular Lasso solutions and computation of all relaxed Lasso solutions is often identically expensive as computing all regular Lasso solutions. Theoretical and numerical results demonstrate that the relaxed Lasso produces sparser models with equal or lower prediction loss than the regular Lasso estimator for high-dimensional data.

Keywords: High dimensionality; Bridge estimation; Lasso; ℓ_q-norm penalisation; Dimensionality reduction

1. Introduction

The current work is motivated by linear prediction for high-dimensional data, where the number of predictor variables p is very large, possibly very much larger than the number of observations n (e.g. van de Geer and van Houwelingen, 2004). Regularisation is clearly of central importance for these high-dimensional problems.

There are many criteria to consider when choosing an appropriate regularisation method. First, not all regularisation procedures are adequate for the high-dimensional case. The non-negative Garotte (Breiman, 1995) is for example a promising regularisation method. However, it is not suited for the case $p > n$ as it requires computation of the OLS-estimator, which is unavailable in this case. An important criterion in the presence of many predictor variables is the computational complexity of the procedure. Many regularisation procedures with otherwise attractive features involve, unfortunately, minimisation of a non-convex function (e.g. Fan and Li, 2001; Tsybakov and van de Geer, 2005). For high-dimensional problems, it is in general very costly to find an (approximate) solution in this case, due to the presence of local minima in the objective function.

For Bridge estimators, which were proposed in Frank and Friedman (1993), we study in the following the tradeoff between computational complexity on the one hand and (asymptotic) properties of the estimators on the other hand. Let $X = (X^1, \ldots, X^p)$ be a p-dimensional predictor variable and Y a response variable of interest. For n independent

E-mail address: nicolai@stat.berkeley.edu.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.019

observations (Y_i, X_i), $i = 1, \ldots, n$, of (Y, X), Bridge estimators are defined for $\lambda, \gamma \in [0, \infty)$ as

$$\hat{\beta}^{\lambda,\gamma} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \|\beta\|_{\gamma}, \tag{1}$$

where $\|\beta\|_{\gamma} = \sum_{k\in\{1,\ldots,p\}} |\beta_k|^{\gamma}$ is the ℓ_{γ}-norm of the vector of coefficients, and γ is typically in the range $[0, 2]$.

For $\gamma = 0$, Bridge estimation corresponds to ordinary model selection. Ridge regression is obtained for $\gamma = 2$, while $\gamma = 1$ is equivalent to the Lasso proposed in Tibshirani (1996). Computation of the estimator (1) involves minimisation of a non-convex function if $\gamma < 1$, while the function is convex for $\gamma \geqslant 1$. Since optimisation of a non-convex function in a high-dimensional setting is very difficult, Bridge estimation with $\gamma \geqslant 1$ is an attractive choice. However, for values of $\gamma > 1$, the shrinkage of estimators towards zero increases with the magnitude of the parameter being estimated (Knight and Fu, 2000). For the Lasso ($\gamma = 1$), the shrinkage is constant irrespective of the magnitude of the parameter being estimated (at least for orthogonal designs, where regularisation with the Lasso is equivalent to soft-thresholding of the estimates). It was recognised in Fan and Li (2001) that this leads to undesirable properties (in terms of prediction) of the resulting estimator. It was first suggested by Huber (1973) to examine the asymptotic properties for a growing number $p = p_n$ of predictor variables as a function of the number of observations n, see as well Fan and Peng (2004). It will be shown below that the shrinkage of the Lasso leads to a low convergence rate of the ℓ_2-loss for high-dimensional problems where the number of parameters $p = p_n$ is growing almost exponentially fast with n, so that $p_n \gg n$.

For $\gamma < 1$, the shrinkage of estimates decreases with increasing magnitude of the parameter being estimated and faster convergence rates can thus in general be achieved (see e.g. Knight and Fu, 2000 and, for classification, Tsybakov and van de Geer, 2005). However, the fact remains that for $\gamma < 1$ a non-convex optimisation problem has to be solved.

There is no value of γ for which an entirely satisfactory compromise is achieved between low computational complexity on the one hand and fast convergence rates on the other hand. In this paper, it is shown that a two-stage procedure, termed relaxed Lasso, can work around this problem. The method has low computational complexity (the computational burden is often identical to that of an ordinary Lasso solution) and, unlike the Lasso, convergence rates are fast, irrespective of the growth rate of the number of predictor variables. Moreover, relaxed Lasso leads to consistent variable selection under a prediction-optimal choice of the penalty parameters, which does not hold true for ordinary Lasso solutions in a high-dimensional setting.

2. Relaxed Lasso

We define relaxed Lasso estimation and illustrate the properties of the relaxed Lasso estimators for an orthogonal design. A two-stage algorithm for computing the relaxed Lasso estimator is then proposed, followed by a few remarks about extending the procedure to generalised linear models (McCullagh and Nelder, 1989).

Recall that the Lasso estimator under a squared error loss is defined in Tibshirani (1996) for $\lambda \in [0, \infty)$ as

$$\hat{\beta}^{\lambda} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \|\beta\|_1. \tag{2}$$

The Lasso estimator is a special case of the Bridge estimator (1), obtained by setting $\gamma = 1$. The set of predictor variables selected by the Lasso estimator $\hat{\beta}^{\lambda}$ is denoted by $\mathcal{M}_{\lambda}$,

$$\mathcal{M}_{\lambda} = \{1 \leqslant k \leqslant p | \hat{\beta}_k^{\lambda} \neq 0\}. \tag{3}$$

For sufficiently large penalties λ (e.g. for $\lambda > 2 \max_k n^{-1} \sum_{i=1}^{n} Y_i X_i^k$), the selected model is the empty set, $\mathcal{M}_{\lambda} = \emptyset$, as all components of the estimator (2) are identical to zero. In the absence of a ℓ_1-penalty and if the number of variables p is smaller than the number of observations n, all predictor variables are in general selected, so that $\mathcal{M}_0 = \{1, \ldots, p\}$ in this case.

The ℓ_1-penalty for the ordinary Lasso estimator (2) has two effects, model selection and shrinkage estimation. On the one hand, a certain set of coefficients is set to zero and hence excluded from the selected model. On the other hand, for all variables in the selected model $\mathcal{M}_{\lambda}$, coefficients are shrunken towards zero compared to the least-squares solution. These two effects are clearly related and can be best understood in the context of orthogonal design as soft-thresholding of the coefficients. Nevertheless, it is not immediately obvious whether it is indeed optimal to control these

two effects, model selection on the one hand and shrinkage estimation on the other hand, by a single parameter only. As an example, it might be desirable in some situations to estimate the coefficients of all selected variables without shrinkage, corresponding to a hard-thresholding of the coefficients.

As a generalisation of both soft- and hard-thresholding, we control model selection and shrinkage estimation by two separate parameters λ and ϕ with the relaxed Lasso estimator.

Definition 1. The relaxed Lasso estimator is defined for $\lambda \in [0, \infty)$ and $\phi \in (0, 1]$ as

$$\hat{\beta}^{\lambda,\phi} = \arg\min_{\beta} n^{-1} \sum_{i=1}^{n} (Y_i - X_i^{\mathrm{T}}\{\beta \cdot \mathbf{1}_{\mathcal{M}_\lambda}\})^2 + \phi\lambda\|\beta\|_1, \tag{4}$$

where $\mathbf{1}_{\mathcal{M}_\lambda}$ is the indicator function on the set of variables $\mathcal{M}_\lambda \subseteq \{1, \ldots, p\}$ so that for all $k \in \{1, \ldots, p\}$,

$$\{\beta \cdot \mathbf{1}_{\mathcal{M}_\lambda}\}_k = \begin{cases} 0, & k \notin \mathcal{M}_\lambda, \\ \beta_k, & k \in \mathcal{M}_\lambda. \end{cases}$$

Note that only predictor variables in the set $\mathcal{M}_\lambda$ are considered for the relaxed Lasso estimator. The parameter λ controls thus the variable selection part, as in ordinary Lasso estimation. The relaxation parameter ϕ controls on the other hand the shrinkage of coefficients. If $\phi = 1$, the Lasso and relaxed Lasso estimators are identical. For $\phi < 1$, the shrinkage of coefficients in the selected model is reduced compared to ordinary Lasso estimation. The case of $\phi = 0$ needs special consideration, as the definition above would produce a degenerate solution. In the following, we define the relaxed Lasso estimator for $\phi = 0$ as the limit of the above definition for $\phi \to 0$. In this case, all coefficients in the model $\mathcal{M}_\lambda$ are estimated by the OLS-solution. This estimator (for $\phi = 0$) was already proposed in Efron et al. (2004) as Lars–OLS hybrid, "using Lars to find the model but not to estimate the coefficients" (Efron et al., 2004). The reduction of the sum of squared residuals of this hybrid method over the ordinary Lasso estimator was found to be small for the studied data set, which contained 10 predictor variables only.

We will show further below that the gains with relaxed Lasso estimation (adaptive ϕ) compared to ordinary Lasso estimation ($\phi = 1$) can be very large. Moreover, relaxed Lasso is producing in most cases better results than the Lars–OLS hybrid ($\phi = 0$), as relaxed Lasso can adapt the amount of shrinkage to the structure of the underlying data.

An algorithm is developed to compute the exact solutions of the relaxed Lasso estimator. The parameters λ and ϕ can then be chosen e.g. by cross-validation. The algorithm is based on the Lars-algorithm by Efron et al. (2004). As the relaxed Lasso estimator is parameterised by two parameters, a two-dimensional manifold has to be covered to find all solutions. The computational burden of computing all relaxed Lasso estimators is in the worst case identical to that of the Lars–OLS hybrid and in the best case identical to that of the Lars-algorithm. The method is thus very well suited for high-dimensional problems.

2.1. Orthogonal design

To illustrate the properties of the relaxed Lasso estimator, it is instructive to consider an orthogonal design. The shrinkage of various regularisation methods are shown in Fig. 1 for this case. The set of solutions of the relaxed Lasso estimator is given for all $k = 1, \ldots, p$ by

$$\hat{\beta}_k^{\lambda,\phi} = \begin{cases} \hat{\beta}_k^0 - \phi\lambda, & \hat{\beta}_k^0 > \lambda, \\ 0, & |\hat{\beta}_k^0| \leqslant \lambda, \\ \hat{\beta}_k^0 + \phi\lambda, & \hat{\beta}_k^0 < -\lambda, \end{cases}$$

where $\hat{\beta}^0$ is the OLS-solution. For $\phi = 0$, hard-thresholding is achieved, while $\phi = 1$ results—as mentioned above—in soft-thresholding, which corresponds to the Lasso solution. The relaxed Lasso provides hence a continuum of solutions that includes soft- and hard-thresholding, much like the set of solutions provided by the Bridge estimators (1) when varying γ in the range $[0, 1]$. It can be seen in Fig. 1 that the solutions to the Bridge estimators and the relaxed Lasso solutions are indeed very similar.

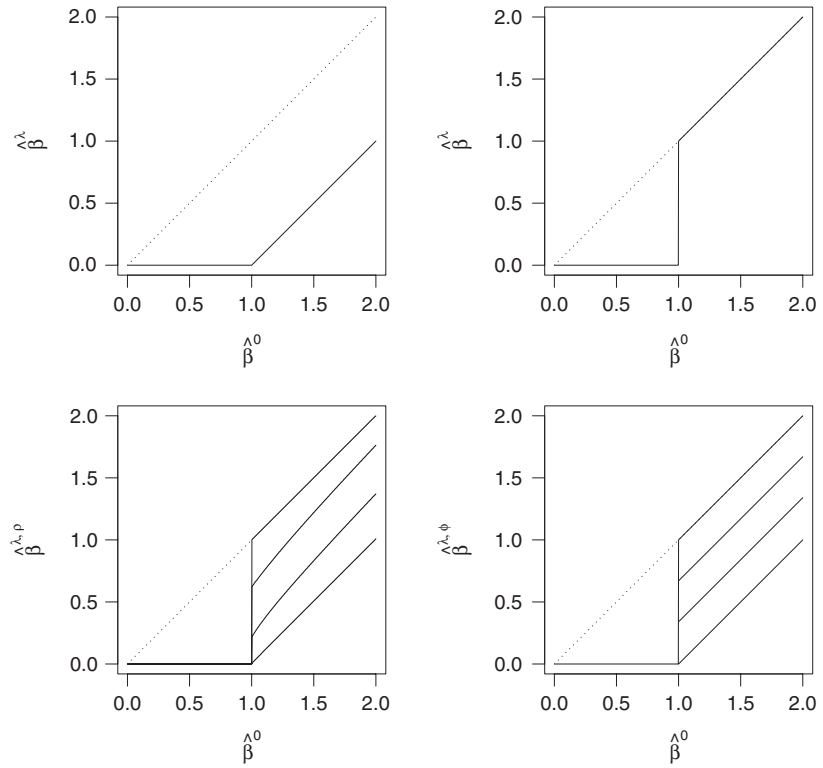

Fig. 1. Comparison of shrinkage estimators as a function of the OLS-estimator $\hat{\beta}^0$. Shown are estimators for soft-thresholding (top left), hard-thresholding (top right), the estimator $\hat{\beta}^{\lambda,\gamma}$, Eq. (1), for $\gamma = 0$, 0.5, 0.9, and 1 (bottom left) and the relaxed Lasso estimators $\hat{\beta}^{\lambda,\phi}$ for $\phi = 0$, $\frac{1}{3}$, $\frac{2}{3}$, and 1 (bottom right).

2.2. Algorithm

The main advantage of the relaxed Lasso estimator over Bridge estimation is the low computational complexity. We propose in the following a naive, easy to implement, algorithm for computing relaxed Lasso estimators as in (4). Based on some more insight, a modification is proposed further below so that—for many data sets—the computational effort of computing all relaxed Lasso solutions is identical to that of solving the ordinary Lasso solutions.

Simple Algorithm.

Step 1: Compute all ordinary Lasso solutions e.g. with the Lars-algorithm in Efron et al. (2004) under the Lasso modification. Let $\mathcal{M}_1, \ldots, \mathcal{M}_m$ be the resulting set of m models. Let $\lambda_1 > \cdots > \lambda_m = 0$ be a sequence of penalty values so that $\mathcal{M}_\lambda = \mathcal{M}_k$ if and only if $\lambda \in (\lambda_k, \lambda_{k-1}]$, where $\lambda_0 := \infty$. (The models are not necessarily distinct, so it is always possible to obtain such a sequence of penalty parameters.)

Step 2: For each $k = 1, \ldots, m$, compute all Lasso solutions on the set $\mathcal{M}_k$ of variables, varying the penalty parameter between 0 and λ_k. The obtained set of solutions is identical to the set of relaxed Lasso solutions $\hat{\beta}^{\lambda,\phi}$ for $\lambda \in \Lambda_k$. The relaxed Lasso solutions for all penalty parameters are given by the union of these sets.

It is obvious that this algorithm produces all relaxed Lasso solutions, for all values of the penalty parameters $\phi \in [0, 1]$ and $\lambda > 0$. The computational complexity of this algorithm is identical to that of Lars–OLS hybrid, as the Lars iterations in Step 2 are about as computationally intensive as ordinary least-squares estimation (Efron et al., 2004).

However, this naive algorithm is not optimal in general. The computation of the ordinary Lasso solutions contains information that can be exploited in the second stage, when finding Lasso solutions for all subsets $\mathcal{M}_k$, $k = 1, \ldots, m$ of variables. Fig. 2 serves as an illustration. The "direction" in which relaxed Lasso solutions are found is identical to the directions of ordinary Lasso solutions. These directions do not have to be computed again. Indeed, by extrapolating the path of the ordinary Lasso solutions, all relaxed Lasso solutions can often be found. There is an important caveat.

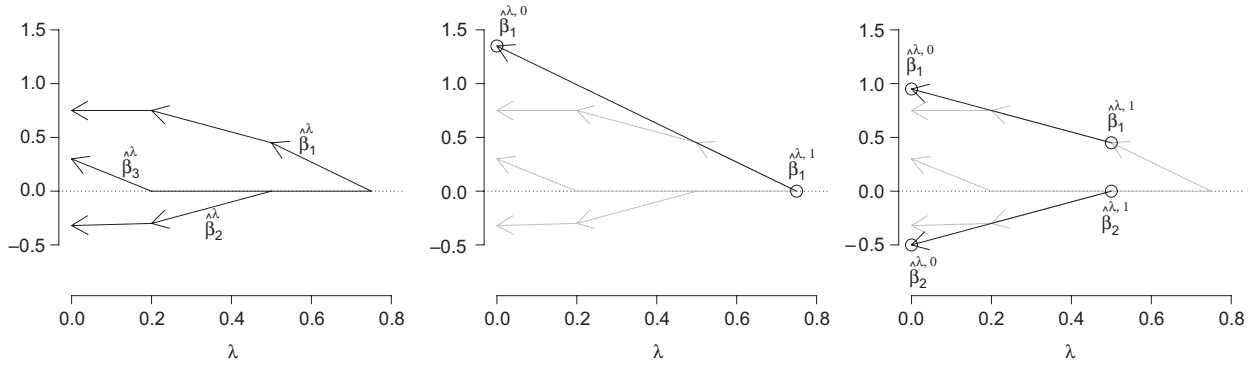

Fig. 2. Left: path of the estimators $\hat{\beta}^{\lambda}$ for a data set with three variables. For large values of λ all components are equal to zero. In the range $\lambda \in (0.45, 0.75]$, only the first components is non-zero. Middle: the relaxed Lasso solutions, if λ is in the range $\lambda \in (0.45, 0.75]$. The direction in which the relaxed Lasso solutions are found is the same as those computed for the ordinary Lasso solutions. The relaxed Lasso solution for $\phi = 0$ corresponds to the OLS-solution. Right: likewise, relaxed Lasso solutions for the range $\lambda \in (0.2, 0.45]$ are found by extrapolating the Lasso solutions. Again, the solutions for $\phi = 0$ correspond to the OLS-solution for the two variables selected by the Lasso estimator.

Extrapolated Lasso solutions are only valid relaxed Lasso solutions if and only if the extrapolations do not cross the value zero. This is e.g. fulfilled if the ordinary Lasso estimators are monotonously increasing for a decreasing penalty parameter λ. If, however, the extrapolations do cross zero for a set $\mathcal{M}_k$, then the Lasso has to be computed again explicitly for this set, using e.g. again the Lars-algorithm of Efron et al. (2004).

Refined Algorithm.

Step 1: Identical to Step 1 of the simple algorithm. Compute all ordinary Lasso solutions.

Step 2: For each $k = 1, \ldots, m$, let $\delta(k) = (\hat{\beta}^{\lambda_k} - \hat{\beta}^{\lambda_{k-1}})/(\lambda_{k-1} - \lambda_k)$. This is the direction in which solutions are found for ordinary Lasso solutions and is hence known from Step 1. Let $\tilde{\beta} = \hat{\beta}^{\lambda_k} + \lambda_k \delta(k)$. If there is at least one component l so that $\text{sign}(\tilde{\beta}_l) \neq \text{sign}(\hat{\beta}_l^{\lambda_k})$, then relaxed Lasso solutions for $\lambda \in \Lambda_k$ have to be computed as in Step 2 of the simple algorithm. Otherwise all relaxed Lasso solutions for $\lambda \in \Lambda_k$ and $\phi \in [0, 1]$ are given by linear interpolation between $\hat{\beta}^{\lambda_{k-1}}$ (which corresponds to $\phi = 1$) and $\tilde{\beta}$ (which corresponds to $\phi = 0$).

In the worst case, the refined algorithm is no improvement over the simple algorithm. In the best case, all relaxed Lasso solutions are found at no extra cost, once the ordinary Lasso solutions are computed. If Lasso solutions are e.g. monotonously increasing (for a decreasing value of λ), then the condition about sign-equality in Step 2 of the refined algorithm is fulfilled, and the relaxed Lasso solutions are found at no extra cost.

The computational complexity of the ordinary Lasso is $O(np \min\{n, p\})$, as there are $m = O(\min\{n, p\})$ steps, each of complexity $O(np)$. In the worst case, the computational complexity of the relaxed Lasso is $O(m^2 np)$, which is, for high-dimensional problems with $p > n$, identical to $O(n^3 p)$, and hence slightly more expensive than the $O(n^2 p)$ of the ordinary Lasso (but equally expensive as the Lars–OLS hybrid if the least-squares estimator is computed explicitly). However, the linear scaling with the number p of variables is identical. Moreover, as mentioned above, the scaling $O(n^3 p)$ is really a worst case scenario. Often all relaxed Lasso solutions can be found at little or no extra cost compared to the ordinary Lasso solutions, using the refined algorithm above.

2.3. Extensions

The method can be easily generalised to more general loss functions and generalised linear models (McCullagh and Nelder, 1989). Let $\ell(\beta)$ be the negative log-likelihood under parameter β. The relaxed Lasso estimator is then defined in analogy to (4) as

$$\hat{\beta}^{\lambda,\phi} = \arg \min_{\beta \in \mathcal{M}_\lambda} \ell(\beta) + \phi\lambda\|\beta\|_1, \tag{5}$$

where $\beta \in \mathcal{M}_\lambda$ is understood to be equivalent to requiring that $\beta_k = 0$, for all $k \notin \mathcal{M}_\lambda$. The algorithm for computing the solutions for all parameter values λ, ϕ has the same two-stage characteristic as for the quadratic loss function.

The computational effort is again identical to that of ordinary Lasso estimation. For this case, no exact solutions for ordinary Lasso estimators are in general available, and the same is true for the relaxed Lasso estimators. However, only optimisation of convex functions are required as long as the log-likelihood is a concave function. For the Lasso, a solution has been proposed e.g. in Zhao and Yu (2004) and could be generalised to compute all relaxed Lasso solutions.

3. Asymptotic results

For the asymptotic results, we consider a random design. Let

$$X = (X^1, \ldots, X^p)$$

be a $p = p_n$-dimensional random variable with a gaussian distribution with covariance matrix Σ, so that $X \sim \mathcal{N}(0, \Sigma)$. The response variable Y is a linear combination of the predictor variables,

$$Y = X^{\mathrm{T}}\beta + \varepsilon, \tag{6}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We compare the risk of the Lasso estimator and the relaxed Lasso estimator. The minimal achievable squared error loss is given by the variance σ^2 of the noise term. The random loss $L(\lambda)$ of the Lasso estimator is defined by

$$L(\lambda) = E(Y - X^{\mathrm{T}}\hat{\beta}^{\lambda})^2 - \sigma^2, \tag{7}$$

where the expectation is with respect to a sample that is independent of the sample which is used to determine the estimator. The loss $L(\lambda, \phi)$ of the relaxed Lasso estimator under parameters λ, ϕ is defined analogously as

$$L(\lambda, \phi) = E(Y - X^{\mathrm{T}}\hat{\beta}^{\lambda,\phi})^2 - \sigma^2. \tag{8}$$

It is shown in the following that convergence rates for the relaxed Lasso estimator are largely unaffected by the number of predictor variables for sparse high-dimensional data. This is in contrast to the ordinary Lasso estimator, where the convergence rate drops dramatically for large growth rates of the number p_n of predictor variables.

3.1. Setting and assumptions

We make a few assumptions about sparse high-dimensional data. The number of predictor variables $p = p_n$ is assumed to be growing very fast with the number of observations.

Assumption 2. For some $c > 0$ and $0 < \xi < 1$,

$$\log p_n \sim cn^{\xi}. \tag{9}$$

In the following, a matrix is said to be diagonally dominant at value v if the row-wise sum of the absolute values of its non-diagonal entries are bounded by v times the corresponding absolute value of the diagonal element.

Assumption 3. There exists some $v < 1$ so that both Σ and Σ^{-1} are diagonally dominant at value v for all $n \in \mathbb{N}$.

Note that a diagonal dominant matrix (for any value $v > 0$) is positive definite. The existence of Σ^{-1} is hence already implied by the assumption about Σ. The assumption is not of critical importance for the results, but shortens the proofs considerably.

The coefficient vector β is assumed to be sparse. For simplicity of exposition, we assume sparseness in the ℓ_0-norm: there are a finite number q of non-zero components of β and these are fix for all $n \in \mathbb{N}$. W.l.o.g., the non-zero components are first in order.

Assumption 4. The vector $\beta \in \mathbb{R}^{p_n}$ of coefficients is given for all $n \in \mathbb{N}$ by $\beta = (\beta_1, \ldots, \beta_q, 0, 0, \ldots)$.

The true model is hence $\mathcal{M}_\star = \{1, \ldots, q\}$. The $p_n - q$ noise variables with zero coefficients are nevertheless possibly correlated with the response variable. This setting is similar to some numerical examples in Fan and Peng (2004).

As the number of non-zero coefficients is given by a finite and fixed number q, we restrict the penalty parameter λ in the following to the range Λ, for which the number of selected variables is less than or equal to $d \log n$ with an arbitrary large $d > 0$,

$$\Lambda := \{\lambda \geqslant 0 : \#\mathcal{M}_\lambda \leqslant d \log n\}. \tag{10}$$

This range includes all sequences λ_n for which the Lasso or relaxed Lasso estimates are consistent for variable selection, as the number of true non-zero coefficients is finite and fixed.

3.2. Slow rates with the ordinary Lasso

It is shown that the rate of convergence of ordinary Lasso estimators is slow if the number of noise variables is growing fast.

Theorem 5. *Under Assumptions* 2–4 *and independent predictor variables, that is* $\Sigma = \mathbf{1}$, *it holds for the risk under the ordinary Lasso estimator that for any* $c > 0$ *and* $n \to \infty$

$$P\left(\inf_{\lambda \in \Lambda} L(\lambda) > cn^{-r}\right) \to 1 \quad \forall r > 1 - \xi.$$

A proof is given in the appendix.

It is hence shown that the rate of convergence of the risk is critically determined by the rate n^{ξ} with which the logarithm $\log p_n$ of the number of predictor variables is growing to infinity. It follows that it is impossible to have both consistent variable selection and optimal rates for independent predictor variables with the ordinary Lasso estimator.

Adding many noise predictor variables slows down the rate of convergence for the Lasso estimator, no matter how the penalty parameter λ is chosen. The reason for this slow convergence in the high-dimensional setting is that a large value of the penalty parameter λ is necessary to keep the estimates of coefficients of noise predictor variables at low values. The shrinkage of the non-zero components is then very large, leading to less than optimal prediction; for a further discussion of this phenomenon see as well Fan and Li (2001).

3.3. Fast rates with the relaxed Lasso

A faster rate of convergence is achieved with the relaxed Lasso estimator than with ordinary Lasso in this sparse high-dimensional setting. Noise variables can be prevented from entering the estimator with a high value of the penalty parameter λ, while the coefficients of selected variables can be estimated at the usual $\sqrt{n}$-rate, using a relaxed penalty. It is shown in other words that the rate of convergence of the relaxed Lasso estimator is not influenced by the presence of many noise variables.

Theorem 6. *Under Assumptions* 2–4, *for* $n \to \infty$, *it holds for the loss under the relaxed Lasso estimator that*

$$\inf_{\lambda \in \Lambda, \phi \in [0,1]} L(\lambda, \phi) = \mathrm{O_p}(n^{-1}).$$

A proof is given in the appendix.

The rate of convergence of the relaxed Lasso estimator (under oracle choices of the penalty parameters) is thus shown to be uninfluenced by a fast growing number of noise variables. The results are illustrated in Fig. 3.

3.4. Choice of the penalty parameters by cross-validation

It was shown above that the rate of convergence of the relaxed Lasso estimate is not influenced by the presence of many noise variables under an oracle choice of the penalty parameters λ and ϕ (which are unknown). We show that the parameters λ, ϕ can be chosen by cross-validation while still retaining the fast rate.

For K-fold cross-validation, each observation belongs to one of K partitions, each consisting of $\tilde{n}$ observations, where $\tilde{n}/n \to 1/K$ for $n \to \infty$. Let $L_{S,\tilde{n}}(\lambda, \phi)$ be for $S = 1, \ldots, K$ the empirical loss on the observations in partition S

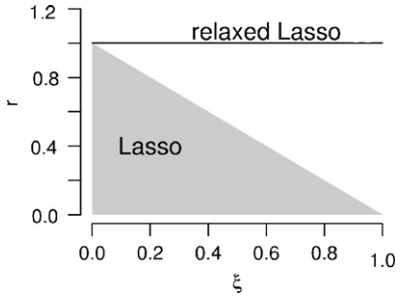

Fig. 3. Convergence rates for the Lasso and the relaxed Lasso. The parameter ξ determines the rate at which the number p_n of variables grows for n, between constant ($\xi = 0$) and exponential ($\xi = 1$). The loss under the relaxed Lasso is $O_p(n^{-1})$, irrespective of ξ. The loss under the ordinary Lasso estimator can be of oder $O_p(n^{-r})$ only if $r < 1 - \xi$ (depicted by the grey area in the figure), no matter how the penalty parameter λ is chosen.

when constructing the estimator on the set of observations different from S. Let $L_{cv}(\lambda, \phi)$ be the empirical loss under K-fold cross-validation,

$$L_{cv}(\lambda, \phi) = K^{-1} \sum_{S=1}^{K} L_{S,\tilde{n}}(\lambda, \phi).$$

The penalty parameters $\hat{\lambda}$ and $\hat{\phi}$ are chosen as minimisers of $L_{cv}(\lambda, \phi)$,

$$(\hat{\lambda}, \hat{\phi}) = \arg \min_{(\lambda,\phi)\in\Lambda\times[0,1]} L_{cv}(\lambda, \phi).$$

In practice, a value of K between 5 and 10 is recommended, even though the following result is valid for a broader range.

Theorem 7. *Let $L(\hat{\lambda}, \hat{\phi})$ be the loss of the relaxed Lasso estimate and $(\hat{\lambda}, \hat{\phi})$ chosen by K-fold cross-validation with $2 \leqslant K < \infty$. Under Assumptions* 2–4, *it holds that*

$$L(\hat{\lambda}, \hat{\phi}) = O_p(n^{-1} \log^2 n).$$

The optimal rates under oracle choices of the penalty parameters are thus almost obtained if the penalty parameters are chosen by cross-validation. We conjecture that the cross-validated penalty parameters lead for the relaxed Lasso estimator to consistent variable selection; this is not the case for the Lasso, see Meinshausen and Bühlmann (2006). This conjecture is supported by the following numerical examples.

4. Numerical examples

We illustrate the asymptotic results of Section 3 with a few numerical examples. The response variable follows the linear model (6). The predictor variable X follows a normal distribution with covariance matrix Σ, where $\Sigma_{ij} = \rho^{|i-j|}$ for some value of $0 \leqslant \rho < 1$. For $\rho = 0$, this corresponds to independent predictor variables. The variance of ε in (6) is chosen so that the signal-to-noise ratio is $0 < \eta < 1$ (e.g. the variance of the response variable Y due to ε is $1/\eta$ of the variance of Y due to $X^T\beta$).

We consider the case where there are q variables (with $q \leqslant p$) that "carry signal" in the sense that $\beta_k \neq 0$ for all $k \leqslant q$ and $\beta_k = 0$ for all k with $k > q$. All components β_k with $k \leqslant q$ are double-exponentially distributed.

For various values of n between 50 and 200 and p between 50 and 800, the ordinary Lasso estimator ($\phi = 1$), the Lars–OLS hybrid estimator ($\phi = 0$), and the relaxed Lasso estimator (adaptive ϕ) are computed. The penalty parameters are chosen by 5-fold cross-validation. The signal-to-noise ratio is chosen from the set $\eta \in \{0.2, 0.8\}$. The correlation between predictor variables is chosen once as $\rho = 0$ (independent predictor variables) and once as $\rho = 0.3$, while the number of relevant predictor variables is chosen from the set $q \in \{5, 15, 25, 50\}$. For each of these settings, the

three mentioned estimators are computed 100 times each. Let L_{rel} be the average loss of relaxed Lasso over these 100 simulations, in the sense of (7), and likewise L_{ols} and L_{lasso} for Lars–OLS hybrid and Lasso, see (8).

For small q, the setting is resembling that of the theoretical considerations above and of the numerical examples in Fan and Peng (2004). For small q, the theorems above suggest that the relaxed Lasso and Lars–OLS hybrid outperform Lasso estimation in terms of predictive power. On the other hand, for $q = p$, the Lasso is the ML estimator and one expects it to do very well compared with the Lars–OLS hybrid for large values of q. (In a Bayesian setting, if the prior for β is chosen to be the actual double-exponential distribution of the components of β, the Lasso solution is the MAP estimator if $q = p$.)

If we knew beforehand the value of q (the number of relevant predictor variables), then we would for optimal prediction either choose Lasso (if q is large), that is $\phi = 1$, or Lars–OLS hybrid (if q is small), that is $\phi = 0$. However, we do not know the value of q. The numerical examples illustrate how well relaxed Lasso adapts to the unknown sparsity of the underlying data.

4.1. Number of selected variables

The number of selected variables is shown in Table 1 for the Lasso estimator and in Table 2 for the relaxed Lasso. As expected, the relaxed Lasso selects roughly the correct number of variables (or less, if the noise is high or the number

Table 1
Average number of selected variables with the *Lasso* for $\rho = 0$

p	50	100	200	400	800	50	100	200	400	800
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	17	18	20	22	25	9	9	8	9	8
100	13	23	24	27	31	10	12	15	14	17
200	11	13	27	31	34	11	13	19	22	24
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	27	30	30	27	24	10	8	8	6	7
100	26	39	44	52	53	13	15	14	16	17
200	27	35	51	59	69	18	21	26	29	30
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	36	30	23	16	12	9	7	8	7	6
100	47	65	66	61	54	15	19	16	14	11
200	48	71	96	112	121	27	30	34	31	27

Table 2
Average number of selected variables with the *relaxed Lasso* for $\rho = 0$

p	50	100	200	400	800	50	100	200	400	800
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	7	6	7	5	5	6	5	6	6	4
100	5	6	6	5	5	6	8	7	6	9
200	5	4	6	5	5	5	6	12	8	8
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	19	18	18	15	12	8	6	6	6	6
100	17	19	16	17	16	10	12	10	9	12
200	15	15	16	15	13	12	14	18	18	15
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	34	25	19	12	9	8	6	6	6	5
100	44	57	55	45	34	13	17	13	11	9
200	46	57	64	71	66	23	26	29	24	19

Table 3
Average number of selected variables with the *Lars–OLS hybrid* for $\rho = 0$

p	50	100	200	400	800	50	100	200	400	800
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	5	5	6	5	5	3	3	2	4	2
100	4	5	5	4	5	4	4	3	3	5
200	4	4	5	4	4	3	3	7	4	4
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	17	16	17	13	11	3	3	3	3	3
100	14	16	16	16	15	5	6	4	4	7
200	13	13	14	14	13	7	6	8	6	4
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	31	23	16	12	7	3	3	2	2	3
100	42	49	48	37	30	5	5	2	5	1
200	46	47	52	58	57	16	10	9	5	4

Table 4
The relative improvement of *relaxed Lasso* over ordinary *Lasso* for $\rho = 0$ (upper half) and $\rho = 0.3$ (lower half)

p	50	100	200	400	800	50	100	200	400	800
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	41	55	49	49	52	−1	0	−2	−2	1
100	95	88	89	110	146	5	6	5	11	9
200	106	88	84	169	171	24	14	11	21	22
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	−5	−3	−2	−3	−2	−3	−3	−4	−4	−4
100	7	17	18	18	12	−3	0	−1	3	−1
200	28	32	43	58	60	−3	2	2	3	4
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	−3	−4	−3	−2	−3	−4	−3	−3	−4	−3
100	−3	−2	−4	−4	−3	−3	−1	−1	−1	−1
200	0	−1	3	1	−2	−4	−1	−1	−1	0
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	40	98	104	85	103	−2	2	0	−3	1
100	83	95	114	180	186	9	8	6	10	15
200	119	128	89	166	202	24	32	10	19	41
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	−3	3	5	−1	7	−3	−5	−3	−2	−2
100	14	31	36	33	49	0	−1	−3	1	0
200	50	48	72	77	114	−4	−1	5	7	3
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	−7	−2	−3	−2	−3	−4	−4	−2	−2	−4
100	−2	−1	0	−2	−3	−3	−1	−1	−3	−1
200	2	7	8	12	11	−2	−3	−1	−2	−1

of observations n is low, with the Lars–OLS hybrid selecting even fewer variables in these cases, as can be seen from Table 3). In contrast, ordinary Lasso often selects too many noise variables (with the cross-validated choice of λ). For $q = 5$, it selects e.g. up to 34 variables. For $q = 50$, up to 121. Using the considerations in the proof of Theorem 5, these numbers can be expected to grow even higher if a larger number n of observations would be considered.

Table 5
The relative improvement of *relaxed Lasso* over *Lars–OLS hybrid* for $\rho = 0$ (upper half) and $\rho = 0.3$ (lower half)

p	50	100	200	400	800	50	100	200	400	800
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	0	3	2	−3	−2	38	29	21	28	20
100	−4	6	1	1	5	18	46	34	19	23
200	8	1	−4	5	0	2	−4	55	5	9
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	8	5	2	2	3	28	21	20	9	9
100	6	3	0	0	0	32	42	31	26	29
200	3	−2	2	−2	0	22	20	48	23	22
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	18	8	7	7	6	20	14	9	3	4
100	6	11	6	6	6	35	33	16	20	8
200	3	6	3	2	2	41	46	52	31	30
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	−3	2	−1	3	−1	42	33	24	30	26
100	−6	−4	−7	−1	3	11	39	20	21	22
200	−3	−1	−2	2	2	1	−4	81	4	−6
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	0	0	−1	1	1	37	29	23	18	9
100	0	2	−2	1	1	29	39	39	33	32
200	−2	2	1	−1	−1	14	13	34	9	16
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	18	1	5	4	5	29	17	10	5	3
100	5	9	8	4	3	46	43	28	17	13
200	4	1	2	1	0	39	36	46	41	31

Table 6
The average value of $\hat{\phi}$ for the *relaxed Lasso*, for $\rho = 0$

p	50	100	200	400	800	50	100	200	400	800
n	$q = 5,\ \eta = 0.8$					$q = 5,\ \eta = 0.2$				
50	0.14	0.09	0.08	0.04	0.03	0.66	0.50	0.49	0.51	0.46
100	0.09	0.11	0.08	0.03	0.04	0.52	0.68	0.53	0.45	0.51
200	0.08	0.08	0.07	0.06	0.06	0.29	0.39	0.71	0.47	0.41
n	$q = 15,\ \eta = 0.8$					$q = 15,\ \eta = 0.2$				
50	0.24	0.15	0.13	0.19	0.20	0.67	0.50	0.45	0.47	0.45
100	0.21	0.17	0.05	0.05	0.04	0.66	0.72	0.61	0.61	0.64
200	0.17	0.12	0.10	0.06	0.02	0.54	0.63	0.75	0.70	0.61
n	$q = 50,\ \eta = 0.8$					$q = 50,\ \eta = 0.2$				
50	0.55	0.38	0.39	0.45	0.43	0.65	0.47	0.45	0.41	0.33
100	0.44	0.54	0.45	0.41	0.40	0.72	0.77	0.71	0.64	0.58
200	0.40	0.44	0.30	0.29	0.19	0.77	0.84	0.89	0.79	0.75

4.2. Comparison with Lasso

Lasso and relaxed Lasso estimators produce nearly identical results (in terms of predictive power) if the number q of relevant predictor variables is large, as can be seen from Table 4, which shows the relative improvement of relaxed

Lasso over ordinary Lasso,

$$100 \cdot (L_{\text{lasso}}/L_{\text{rel}} - 1). \tag{11}$$

There is no harm when using the relaxed Lasso on such data instead of the Lasso, but there is not much to be gained either. However, for data where there is a very large number of noise variables (e.g. small q), the relaxed Lasso estimator produces a much smaller MSE, as expected from the previous theoretical results. The extent to which the relaxed Lasso outperforms Lasso in this setting depends strongly on the signal-to-noise ratio η. The improvements are larger for large η, where shrinkage of the selected components is not necessary. For small η, shrinkage of the selected components is useful and an optimal procedure chooses thus ϕ close to 1 for noisy problems. Indeed, the average chosen value of ϕ for the relaxed Lasso is large if η is low, as can be seen from Table 6.

In the worst case, relaxed Lasso is performing only marginally worse than ordinary Lasso and is slightly more expensive to compute. For many sparse high-dimensional problems, however, the computation of the relaxed Lasso solutions comes at no extra computational cost and leads to sparser estimators and more accurate predictions.

4.3. Comparison with Lars–OLS hybrid

The theoretical conclusions suggest that Lars–OLS hybrid should do equally well for sparse high-dimensional data as relaxed Lasso. However, there are two caveats. First, the argument holds only for data with sparse structure. If the data do not have sparse structure, Lars–OLS hybrid is in general performing worse than Lasso. Relaxed Lasso can adapt to the amount of sparseness (as seen from Table 6) by varying ϕ between 1 (for not so sparse data) to 0 (for sparse data). Table 5 shows the relative improvement of relaxed Lasso over Lars–OLS hybrid, analogously to (11). For large values of q, relaxed Lasso is indeed performing better than Lars–OLS in general.

What is more striking than the dependence on the sparseness, however, is the dependence on the signal-to-noise ratio. Consider the case where only five variables carry signal ($q = 5$). For a high signal-to-noise ratio ($\eta = 0.8$), relaxed Lasso and Lars–OLS hybrid perform approximately equally well (and both much better than ordinary Lasso). For a low signal-to-noise ratio ($\eta = 0.2$), however, relaxed Lasso is considerably better than Lars–OLS. The reason for this is intuitively easy to understand. For noisy problems, it pays off to shrink the coefficients of selected variables, while this is less important for less noisy data. Relaxed Lasso adapts the amount of shrinkage to the noise level.

In general, it is not optimal to do no shrinkage at all for the selected variables ($\phi = 0$) or do full shrinkage ($\phi = 1$). This is the reason why relaxed Lasso is performing better than both ordinary Lasso and Lars–OLS hybrid for noisy problems, especially when just a few variables carry signal. Given that the computational cost of relaxed Lasso is not higher than that for Lars–OLS hybrid (and sometimes equal to that of Lasso), relaxed Lasso seems to be well suited for high-dimensional problems as the sparseness and signal-to-noise ratio is in general unknown and relaxed Lasso is adaptive to both (Table 6).

5. Conclusions

We have proposed the relaxed Lasso as a generalisation of Lasso estimation. The main motivation are very high-dimensional regression problems, where the Lasso has two shortcomings:

- *Selection of noise variables:* If the penalty parameter is chosen by cross-validation, the number of selected variables is often very large. Many noise variables are potentially selected.
- *Low accuracy of predictions:* The accuracy of prediction (in terms of squared error loss) was shown to be negatively affected by the presence of many noise variables, particularly for high signal-to-noise ratios.

The advantages of relaxed Lasso over ordinary Lasso in this high-dimensional setting are twofold.

- *Sparser estimates:* The number of selected coefficients is in general very much smaller for relaxed Lasso, without compromising on the accuracy of predictions. The models produced by relaxed Lasso are thus more amenable to interpretation.
- *More accurate predictions:* If the signal-to-noise ratio is very low, the predictive accuracy of both Lasso and relaxed Lasso is comparable. For a high signal-to-noise ratio, relaxed Lasso achieves often much more accurate predictions.

For high signal-to-noise ratios, both advantages of relaxed Lasso—sparser estimates and more accurate predictions—can be achieved alternatively by using the Lars–OLS hybrid. However, Lars–OLS hybrid is not adaptive to the signal-to-noise ratio, as seen in the numerical examples and is performing very much worse than ordinary Lasso for low signal-to-noise ratios. Relaxed Lasso is adaptive to the signal-to-noise ratio and achieves near-optimal performance on a wide variety of data sets.

6. Proofs

6.1. Proof of Theorem 6

It was assumed that the set of non-zero coefficients of β is given by $\mathcal{M}_\star = \{1, \ldots, q\}$. Denote by $\mathcal{E}$ the event

$$\exists \lambda: \ \mathcal{M}_\lambda = \mathcal{M}_\star. \tag{12}$$

Let $c > 0$ be any positive constant. Then

$$P\left(\inf_{\lambda,\phi} L(\lambda, \phi) > cn^{-1}\right) \leqslant P\left(\inf_{\lambda,\phi} L(\lambda, \phi) > cn^{-1} | \mathcal{E}\right) P(\mathcal{E}) + P(\mathcal{E}^c).$$

Let $\lambda_\star$ be the smallest value of the penalty parameter λ such that no noise variable enters the selected variables, that is $\hat{\beta}_k^\lambda = 0$ for all $k > q$,

$$\lambda_\star := \min_{\lambda \geqslant 0} \{\lambda | \hat{\beta}_k^\lambda = 0, \forall k > q\}. \tag{13}$$

The loss $\inf_{\lambda,\phi} L(\lambda, \phi)$ is smaller than $L(\lambda_\star, 0)$. Note that, conditional on $\mathcal{E}$, the loss $L(\lambda_\star, 0)$ is the loss of the regular OLS-estimator $\hat{\beta}^{\star 0}$ on the set $\mathcal{M}_\star = \{1, \ldots, q\}$ of the q predictor variables with non-vanishing coefficients. Let $L^\star$ be the loss of this OLS-estimator. It follows that

$$P\left(\inf_{\lambda,\phi} L(\lambda, \phi) > cn^{-1}\right) \leqslant P(L^\star > cn^{-1} | \mathcal{E}) P(\mathcal{E}) + P(\mathcal{E}^c) \leqslant P(L^\star > cn^{-1}) + P(\mathcal{E}^c).$$

It follows from the proofs in Meinshausen and Bühlmann (2006) that there is a value of λ such that the true model $\mathcal{M}_\star$ is selected with the Lasso estimator, so that $P(\mathcal{E}^c) \to 0$ for $n \to \infty$. By the known properties of the OLS-estimator, there exists some $c > 0$ for every $\varepsilon > 0$, so that $\limsup_{n\to\infty} P(L^\star > cn^{-1}) < \varepsilon$, which completes the proof. □

6.2. Some useful lemmas

6.2.1. Eigenvalues

Let $\Sigma(\mathcal{M})$ be the covariance matrix, restricted to the subset $\mathcal{M} \subseteq \{1, \ldots, p\}$ of variables. Let $\Sigma_n(\mathcal{M})$ be the corresponding empirical covariance matrix for n independent observations.

Lemma 8. *Under Assumptions* 2–4, *there exist* $0 < b_{\min} < b_{\max} < \infty$, *so that the maximal and minimal eigenvalues* $\lambda_{\max}(\mathcal{M})$ *and* $\lambda_{\min}(\mathcal{M})$ *of the empirical covariance matrices* $\Sigma_n(\mathcal{M})$ *are all bounded simultaneously for any* $d > 0$ *and all* $\mathcal{M}$ *with* $|\mathcal{M}| = m_n \leqslant d \log n$ *by* $b_{\min}$ *from below and* $b_{\max}$ *from above, with probability converging to* 1 *for* $n \to \infty$,

$$P(b_{\min} < \lambda_{\min}(\mathcal{M}), \lambda_{\max}(\mathcal{M}) < b_{\max}, \ \forall \mathcal{M} : |\mathcal{M}| \leqslant m_n) \to 1, \quad n \to \infty.$$

Proof. By Gershgorins theorem, all eigenvalues of the empirical covariance matrix $\Sigma_n(\mathcal{M})$ are in the set

$$\Gamma(\mathcal{M}) := \bigcup_{a \in \mathcal{M}} \left\{ x : |x - (\Sigma_n(\mathcal{M}))_{aa}| \leqslant \sum_{b \in \mathcal{M} \setminus a} |(\Sigma_n(\mathcal{M}))_{ab}| \right\}.$$

Let $\Gamma := \{1, \ldots, p\}$ be the set of all predictor variables. Taking the union over all sets with $|\mathcal{M}| \leqslant m_n$,

$$\bigcup_{\mathcal{M}} \Gamma(\mathcal{M}) \subseteq \bigcup_{a \in \{1,\ldots,p\}} \left\{ x : |x - (\Sigma_n)_{aa}| \leqslant \max_{\Xi \subseteq \{1,\ldots,p\}, |\Xi| \leqslant m_n - 1} \sum_{b \in \Xi} |(\Sigma_n)_{ab}| \right\}.$$

Denoting the maximal difference between the covariance matrix and its empirical version by

$$\Delta = \max_{a,b} |(\Sigma_n - \Sigma)_{ab}|, \tag{14}$$

it follows that

$$\bigcup_{\mathcal{M}} \Gamma(\mathcal{M}) \subseteq \bigcup_{a \in \{1,\ldots,p\}} \left\{ x : |x - \Sigma_{aa}| \leqslant m_n \Delta + \sum_{b \neq a} |\Sigma_{ab}| \right\}.$$

Using the assumption that Σ is diagonally dominant at value $v < 1$ and $\Sigma_{aa} = 1$, for all $a \in \{1, \ldots, p\}$, it follows that

$$\bigcup_{\mathcal{M}} \Gamma(\mathcal{M}) \subseteq \bigcup_{a \in \{1,\ldots,p\}} \{x : 1 - v - m_n \Delta < x \leqslant 1 + v + m_n \Delta\}.$$

As $\log p_n \sim cn^{\xi}$ with $\xi < 1$ and $m_n \leqslant d \log n$ for some $d > 0$, it is sufficient to show that there exist $g > 0$ for every $\delta > 0$ so that for $n \to \infty$,

$$P(\Delta \geqslant \delta / m_n) = \mathrm{O}(p_n^2 \exp(-gn/m_n)). \tag{15}$$

Using Bernstein's inequality, there exists $g > 0$ so that for any $1 \leqslant a, b \leqslant p_n$ and for $n \to \infty$,

$$P\left(\left| n^{-1} \sum_{i=1}^{n} (X_i^a X_i^b) - E(X^a X^b) \right| > \delta / m_n \right) = \mathrm{O}(\exp(-gn/m_n)).$$

With Bonferronis inequality, Eq. (15) follows, which completes the proof. □

6.2.2. Change in gradient

Let $\mathcal{V}_h$ be the set of all diagonal $h \times h$ matrices V, where the diagonal elements are in $\{-1, 1\}$.

Lemma 9. *It holds under Assumptions* 2–4 *that, for every* $g > 0$, *with probability converging to* 1 *for* $n \to \infty$, *simultaneously for all* $\mathcal{M}$ *with* $|\mathcal{M}| \leqslant m_n = d \log n$ *and* $V \in \mathcal{V}_{|\mathcal{M}|}$,

$$|\Sigma(\mathcal{M}) \Sigma_n(\mathcal{M})^{-1} V \mathbf{1}_{\mathcal{M}} - V \mathbf{1}_{\mathcal{M}}| < g,$$

where the inequality is understood to be fulfilled if it is fulfilled componentwise.

Proof. First,

$$\Sigma(\mathcal{M}) \Sigma_n(\mathcal{M})^{-1} V \mathbf{1}_{\mathcal{M}} = V \mathbf{1}_{\mathcal{M}} + (\Sigma(\mathcal{M}) - \Sigma_n(\mathcal{M})) \Sigma_n(\mathcal{M})^{-1} V \mathbf{1}_{\mathcal{M}}.$$

Thus, simultaneously for all $\mathcal{M}$ with $|\mathcal{M}| \leqslant m_n$, it holds componentwise that

$$|\Sigma(\mathcal{M}) \Sigma_n(\mathcal{M})^{-1} V \mathbf{1}_{\mathcal{M}} - V \mathbf{1}_{\mathcal{M}}| \leqslant m_n \Delta \max_{\mathcal{M}, a \in \mathcal{M}} |(\Sigma_n(\mathcal{M})^{-1} V \mathbf{1}_{\mathcal{M}})_a|,$$

where Δ is defined as in (14). The last term $\max_{\mathcal{M}, a \in \mathcal{M}} |(\Sigma_n(\mathcal{M})^{-1} V \mathbf{1}_{\mathcal{M}})_a|$ is bounded by $m_n / \lambda_{\min}$, where $\lambda_{\min}$ is the minimal eigenvalue of $\Sigma_n(\mathcal{M})$ over all subsets $\mathcal{M}$ with $|\mathcal{M}| \leqslant m_n$. This minimal eigenvalue is bounded from below by $b_{\min} > 0$ with probability converging to 1 for $n \to \infty$, according to Lemma 8. It remains to be shown that for any $\delta > 0$, $P(\Delta > \delta / m_n^2) \to 1$ for $n \to \infty$. This follows analogously to (15), which completes the proof. □

6.2.3. Restricted positive cone condition

The *positive cone condition* of Efron et al. (2004) is fulfilled if, for all subsets $\mathscr{M} \subseteq \{1, \ldots, p_n\}$ and all $V \in \mathscr{V}_{|\mathscr{M}|}$,

$$(V\Sigma_n(\mathscr{M})V)^{-1}\mathbf{1}_{\mathscr{M}} > 0,$$

where the inequality holds componentwise. The *restricted positive cone condition* is fulfilled if the inequality holds for all subsets $\mathscr{M}$ so that $|\mathscr{M}| \leqslant m_n$.

Lemma 10. *Under Assumptions* 2–4, *the restricted positive cone condition is fulfilled for* $m_n \leqslant d \log n$ *with any* $d > 0$, *with probability converging to* 1 *for* $n \to \infty$. *Moreover, for any* $0 < \varepsilon < 1 - v$,

$$P\left(\min_{\mathscr{M}:|\mathscr{M}|\leqslant m_n, V\in\mathscr{V}_{|\mathscr{M}|}} (V\Sigma_n(\mathscr{M})V)^{-1}\mathbf{1}_{\mathscr{M}} > \varepsilon\right) \to 1, \quad n \to \infty.$$

Proof. First, for any $\mathscr{M}$ and $V \in \mathscr{V}_{|\mathscr{M}|}$,

$$(V\Sigma_n(\mathscr{M})V)^{-1}\mathbf{1}_{\mathscr{M}} = (V\Sigma(\mathscr{M})V)^{-1}(V\Sigma(\mathscr{M})\Sigma_n(\mathscr{M})^{-1}V\mathbf{1}_{\mathscr{M}}).$$

By Lemma 9, the components of $V\Sigma(\mathscr{M})\Sigma_n(\mathscr{M})^{-1}V\mathbf{1}_{\mathscr{M}}$ are, for every $\delta > 0$, simultaneously bounded for all $\mathscr{M}$ with $|\mathscr{M}| \leqslant m_n$ and $V \in \mathscr{V}_{|\mathscr{M}|}$ by $1 - \delta$ from below and by $1 + \delta$ from above, with probability converging to 1 for $n \to \infty$. Thus it holds for every $a \in \mathscr{M}$ and $V \in \mathscr{V}_{|\mathscr{M}|}$, with probability converging to 1 for $n \to \infty$,

$$\begin{aligned}((V\Sigma_n(\mathscr{M})V)^{-1}\mathbf{1}_{\mathscr{M}})_a &\geqslant \Sigma(\mathscr{M})^{-1}_{aa}(1-\delta) - \sum_{b\in\mathscr{M}\setminus a} |\Sigma(\mathscr{M})^{-1}_{ab}|(1+\delta)\\ &= (1-\delta)\left(\Sigma(\mathscr{M})^{-1}_{aa} - \frac{1+\delta}{1-\delta}\sum_{b\in\mathscr{M}\setminus a}|\Sigma(\mathscr{M})^{-1}_{ab}|\right)\\ &=: g_a(\delta).\end{aligned}$$

The inverse covariance matrix Σ^{-1} is by assumption diagonally dominant at value $v < 1$, which is equivalent to

$$\sum_{b\in\mathscr{M}\setminus a} |\Sigma^{-1}_{ab}| \leqslant v\Sigma^{-1}_{aa}.$$

It is straightforward to show that in this case, for all $\mathscr{M} \subseteq \{1, \ldots, p\}$, the inverse covariance matrices $\Sigma(\mathscr{M})^{-1}$ are diagonally dominant at value $v < 1$ as well. For $\delta = 0$, the continuous function $g_a(\delta)$ is hence, for all components $a \in \mathscr{M}$, larger than or equal to $(1 - v)(\Sigma_{aa}(\mathscr{M})^{-1})$. Note that $\Sigma_{aa}(\mathscr{M})^{-1}$ is the inverse of the conditional variance $Var(X^a|\{X^b, b \in \mathscr{M}\setminus a\})$, which is smaller than the unconditional variance $Var(X^a)$. Hence, as $\Sigma_{aa} = 1$, it holds that $\Sigma_{aa}(\mathscr{M})^{-1} > 1$ for all $a \in \mathscr{M}$ and thus for all $a \in \mathscr{M}$,

$$\lim_{\delta\to 0} g_a(\delta) \geqslant 1 - v.$$

Choosing δ sufficiently small, the continuous function $g_a(\delta)$ is for all components $a \in \mathscr{M}$ larger than ε, as $\varepsilon < 1 - v$, which completes the proof. □

Lemma 11. *Under Assumptions* 2–4, *for some* $d > 0$, *it holds for any* $\varepsilon > 1 + v$ *that*

$$P\left(\max_{\mathscr{M}:|\mathscr{M}|\leqslant m_n, V\in\mathscr{V}_{|\mathscr{M}|}} (V\Sigma_n(\mathscr{M})V)^{-1}\mathbf{1}_{\mathscr{M}} < \varepsilon\right) \to 1, \quad n \to \infty.$$

Proof. The proof of Lemma 11 follows analogously to the proof of Lemma 10. □

6.2.4. Monotonicity of Lasso-solutions

Lemma 12. *Under the restricted positive cone condition, the absolute value of the Lasso estimator $\hat{\beta}_k^\lambda$ is for all components $k = 1, \ldots, p$ monotonously increasing for a decreasing value of λ.*

Proof. For any value of λ, let $\delta\lambda > 0$ be a small change of the penalty parameter λ. Let $\delta\hat{\beta}^\lambda$ be the corresponding change of the Lasso estimator,

$$\delta\hat{\beta}^\lambda := \hat{\beta}^{\lambda-\delta\lambda} - \hat{\beta}^\lambda.$$

It has to be shown that for any $\lambda > 0$,

$$\hat{\beta}^\lambda \cdot \delta\hat{\beta}^\lambda \geqslant 0. \tag{16}$$

For all components of $\hat{\beta}^\lambda$ equal to zero, the claim is automatically fulfilled. Let the set of non-zero components of $\hat{\beta}^\lambda$ be again denoted by $\mathcal{M}_\lambda \subseteq \{1, \ldots, p\}$. Denote the restriction of $\hat{\beta}^\lambda$ and $\delta\hat{\beta}^\lambda$ to the set $\mathcal{M}$ by $\hat{\beta}^\lambda(\mathcal{M})$ and $\delta\hat{\beta}^\lambda(\mathcal{M})$, respectively. It follows e.g. from Efron et al. (2004) that the infinitesimal change $\delta\beta(\mathcal{M})$ of the vector $\hat{\beta}^\lambda(\mathcal{M})$ is proportional to

$$(\Sigma_n(\mathcal{M})V)^{-1}\mathbf{1}_{\mathcal{M}}, \tag{17}$$

where V is a diagonal $|\mathcal{M}| \times |\mathcal{M}|$-matrix with diagonal elements V_{kk}, $k \in \mathcal{M}$, identical to the signs of the correlations of X_i^k, $i = 1, \ldots, n$ with the residuals $Y_i - \sum_{a\in\{1,\ldots,p\}} \hat{\beta}_k^\lambda X_i^a$, $i = 1, \ldots, n$. As $\hat{\beta}^\lambda$ is a Lasso solution, V_{kk} is identical to the sign of $\hat{\beta}_k^\lambda$ for all $k \in \mathcal{M}$. Therefore, componentwise, for all $\lambda > 0$

$$\operatorname{sign}(\delta\hat{\beta}^\lambda(\mathcal{M}) \cdot \hat{\beta}^\lambda(\mathcal{M})) = \operatorname{sign}((V\Sigma_n(\mathcal{M})V)^{-1}\mathbf{1}_{\mathcal{M}}).$$

If the restricted positive cone condition is fulfilled, all components on the r.h.s. are positive and so the same is true for the l.h.s., and (16) follows. The restricted positive cone condition is fulfilled with probability converging to 1 for $n \to \infty$ according to Lemma 10, which completes the proof. □

6.2.5. When do noise variables enter?

By assumption, the correct model is given by the first q predictor variables, $\mathcal{M}_\star = \{1, \ldots, q\}$. A noise variable is hence a variable with index larger than q. If any noise variable is part of the Lasso estimator, then, equivalently, there exists some $k > q$ so that $k \in \mathcal{M}_\lambda$.

Lemma 13. *Let λ_n, $n \in \mathbb{N}$, be a sequence with $\lambda_n = o(n^{(-1+\xi)/2})$ for $n \to \infty$. Then, under Assumptions 2–4 and independent predictor variables,*

$$P(\exists k > q : k \in \mathcal{M}_{\lambda_n}) \to 1, \quad n \to \infty.$$

Proof. Let $\hat{\beta}^{\star\lambda}$ be the Lasso estimator, which is constrained to be zero outside the set $\mathcal{M}_\star = \{1, \ldots, q\}$,

$$\hat{\beta}^{\star\lambda} = \arg\min_\beta n^{-1}\sum_{i=1}^n \left(Y_i - \sum_{k\in\mathcal{M}_\star} \beta_k X_i^k\right)^2 + \lambda\|\beta\|_1. \tag{18}$$

If $\hat{\beta}^{\star\lambda}$ is a valid Lasso solution to the unconstrained problem, as in Eq. (2), then there does not exist, by uniqueness of the solution, any $k > q$ so that $k \in \mathcal{M}_\lambda$. It suffices hence to show that $\hat{\beta}^{\star\lambda_n}$ cannot be the solution to (2), with probability converging to 1 for $n \to \infty$. Using results in Osborne et al. (2000), the Lasso estimator $\hat{\beta}^{\star\lambda}$ is only a valid Lasso solution for the whole set of p_n predictor, if the gradient $n^{-1}\sum_{i=1}^n R_i X_i^k$ is smaller or equal to λ for all $k > q$, where, for $i = 1, \ldots, n$,

$$R_i = Y_i - \sum_{a\in\mathcal{M}_\star} \hat{\beta}_k^{\star\lambda} X_i^a,$$

are the residuals under the estimator $\hat{\beta}^{\star\lambda}$. Thus

$$P(\exists k > q : k \in \mathcal{M}_{\lambda_n}) \geqslant P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} R_i X_i^k > \lambda_n\right). \tag{19}$$

Conditional on $(Y, X^1, \ldots, X^q)$, it holds for every $k > q$, that

$$n^{-1} \sum_{i=1}^{n} R_i X_i^k \sim \mathcal{N}\left(0, n^{-2} \sum_{i=1}^{n} R_i^2\right).$$

The expected value of $n^{-1}\sum_{i=1}^{n} R_i^2$, the averaged squared residuals, is larger than $\sigma^2(n-q)/n$ for all values of λ and

$$P\left(n^{-1} \sum_{i=1}^{n} R_i^2 > \sigma^2/2\right) \to 1, \quad n \to \infty.$$

If $n^{-1}\sum_{i=1}^{n} R_i^2 = \sigma^2/2$, then $n^{-1}\sum_{i=1}^{n} R_i X_i^k \sim \mathcal{N}(0, \sigma^2/(2n))$. Thus, for some $c, d > 0$,

$$P\left(n^{-1} \sum_{i=1}^{n} R_i X_i^k > \lambda_n\right) \geqslant d\lambda_n^{-1} \exp(-cn\lambda_n^2),$$

which holds for every $k > q$, of which there are $p_n - q$ variables. The probability that the gradient $n^{-1}\sum_{i=1}^{n} R_i X_i^k$ is smaller than λ_n for all $p_n - q$ noise variables is hence bounded by

$$\begin{aligned} P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} R_i X_i^k \leqslant \lambda_n\right) &\leqslant (1 - d\lambda_n^{-1} \exp(-cn\lambda_n^2))^{p_n - q} \\ &\leqslant \exp(-(p_n - q) d\lambda_n^{-1} \exp(-cn\lambda_n^2)). \end{aligned}$$

Let λ_n be a sequence with $\lambda_n = o(n^{(-1+\xi)/2})$ for $n \to \infty$. Then $n\lambda_n^2 = o(n^{\xi})$, and as $\log p_n \sim g n^{\xi}$, for some $g > 0$, it follows that

$$P\left(\max_{k>q} n^{-1} \sum_{i=1}^{n} R_i X_i^k \leqslant \lambda_n\right) \to 0, \quad n \to \infty,$$

which, using (19), completes the proof. □

6.2.6. Error of estimators

The following lemma bounds from below the difference between the estimator under $\lambda \geqslant \lambda_\star$ and the true parameter value.

Lemma 14. *Assume* $\Sigma = \mathbf{1}$ *and Assumptions* 2 *and* 3. *For any* $\delta > 0$, *with probability converging to* 1 *for* $n \to \infty$, *it holds for all* $k \leqslant q$ *that for* $\lambda \geqslant \lambda_\star$,

$$|\hat{\beta}_k^{\lambda} - \beta_k| \geqslant (1 - \delta)\lambda.$$

Proof. First,

$$|\hat{\beta}_k^{\lambda} - \beta_k| \geqslant |\hat{\beta}_k^{\lambda} - \hat{\beta}_k^{\star 0}| - |\hat{\beta}_k^{\star 0} - \beta_k|,$$

where $\hat{\beta}^{\star 0}$ is defined as in (18) as the Lasso estimator where all components of noise variables, for $k > q$, are restricted to be zero. This estimator is the regular OLS estimator on the set $\mathcal{M} = \{1, \ldots, q\}$ of variables and it follows by standard results that for any series c_n with $c_n^{-1} = o_p(n^{-1/2})$, it holds that $P(|\hat{\beta}_k^{\star 0} - \beta_k| > c_n) \to 0$ for $n \to \infty$. By Lemma 13, $\lambda_\star^{-1} = o_p(n^{-1/2})$. It suffices hence to show that, for any $\delta > 0$, for all $k \leqslant q$ and $\lambda \geqslant \lambda_\star$,

$$P(|\hat{\beta}_k^{\lambda} - \hat{\beta}_k^{\star 0}| > (1 - \delta)\lambda) \to 1, \quad n \to \infty. \tag{20}$$

Note that for $\lambda \geqslant \lambda_\star$, $\hat{\beta}^\lambda = \hat{\beta}^{\star\lambda}$ by definition of $\hat{\beta}^{\star\lambda}$ in (18). Using (17), it holds for every $\lambda > 0$ that

$$|\hat{\beta}_k^{\star\lambda} - \hat{\beta}_k^{\star 0}| = \left| \int_0^\lambda (V \Sigma_n(\mathscr{M}_{\lambda'}) V)^{-1} \mathbf{1}_{\mathscr{M}_{\lambda'}} \, \mathrm{d}\lambda' \right|,$$

where $\mathscr{M}_\lambda = \{k \leqslant q : \hat{\beta}_k^{\star\lambda} \neq 0\} \subseteq \mathscr{M}_\star$. By Lemma 10 and $\Sigma = \mathbf{1}$, it holds for every $\delta > 0$ with probability converging to 1 for $n \to \infty$ that

$$\min_{\mathscr{M}: |\mathscr{M}| \leqslant m_n, V} (V \Sigma_n(\mathscr{M}) V)^{-1} \mathbf{1}_{\mathscr{M}} > (1 - \delta). \tag{21}$$

As $q = |\mathscr{M}_\star| \leqslant m_n$, it follows that with probability converging to 1 for $n \to \infty$,

$$|\hat{\beta}_k^{\star\lambda} - \hat{\beta}_k^{\star 0}| \geqslant (1 - \delta)\lambda,$$

which shows, using $\hat{\beta}^\lambda = \hat{\beta}^{\star\lambda}$ for $\lambda \geqslant \lambda_\star$, that (20) holds true and thus completes the proof. □

6.2.7. Errors due to finite validation set

Let $L_{\tilde{n}}(\lambda, \phi)$ be the empirical version of $L(\lambda, \phi)$ for $\tilde{n}$ observations of (Y, X), which are independent of the observations used to construct the relaxed Lasso estimator.

Lemma 15. *Let* $\liminf_{n\to\infty} \tilde{n}/n \to 1/K$ *with* $K \geqslant 2$. *Then, under Assumptions* 2–4,

$$\sup_{\lambda \in \Lambda, \phi \in [0,1]} |L(\lambda, \phi) - L_{\tilde{n}}(\lambda, \phi)| = \mathrm{O_p}(n^{-1} \log^2 n), \quad n \to \infty.$$

Proof. The restricted positive cone condition is satisfied with probability converging to 1 for $n \to \infty$, according to Lemma 10. It hence suffices to show the claim under assumption of the restricted positive cone condition. Let, as before, $\mathscr{M}_1, \ldots, \mathscr{M}_m$ be the set of all models attained with Lasso estimates and let λ_k, $k = 1, \ldots, m$, (with $\lambda_1 < \cdots < \lambda_m$) be the largest value of the penalty parameter λ so that $\mathscr{M}_k = \mathscr{M}_\lambda$. Using Lemma 12 and the definition of the relaxed Lasso estimates, Eq. (4), any relaxed Lasso solution is in one of the sets $\mathscr{B}_1, \ldots, \mathscr{B}_m$, where for all $k \in \{1, \ldots, m\}$,

$$\mathscr{B}_k = \{\beta = \phi \hat{\beta}^{\lambda_k, 0} + (1 - \phi)\hat{\beta}^{\lambda_k, 1} | \phi \in [0, 1]\}. \tag{22}$$

The estimates $\hat{\beta}^{\lambda_k,1}$ are the Lasso estimates for penalty parameter λ_k, and $\hat{\beta}^{\lambda_k,0}$ the corresponding OLS-estimates. The loss under a choice of λ, ϕ as penalty parameters is given by

$$L(\lambda, \phi) = E\left(Y - \sum_{k \in \{1,\ldots,p\}} \hat{\beta}_k^{\lambda,\phi} X^k\right)^2.$$

For any λ, set $\delta\hat{\beta}^\lambda = (\hat{\beta}^{\lambda,1} - \hat{\beta}^{\lambda,0})$. The loss $L(\lambda, \phi)$ can then be written as

$$L(\lambda, \phi) = E(U_\lambda^2) + 2\phi E(U_\lambda V_\lambda) + \phi^2 E(V_\lambda^2), \tag{23}$$

where $U_\lambda = Y - \sum_{k \in \{1,\ldots,p\}} \hat{\beta}_k^{\lambda,0} X^k$, and $V_\lambda = \sum_{k \in \{1,\ldots,p\}} \delta\hat{\beta}_k^\lambda X^k$. The loss $L(\lambda, \phi)$ is hence, for a given λ, a quadratic function in ϕ. Both U_λ and V_λ are normal distributed random variables conditional on the sample on which $\hat{\beta}^{\lambda,\phi}$ is estimated. There exists some $h > 0$ so that, for all λ and ϕ, $P(\max_k \hat{\beta}_k^{\lambda,\phi} > h) \to 0$ for $n \to \infty$. As the number of non-zero coefficients is bounded by $m_n \leqslant d \log n$, it thus follows by Bernstein's inequality that there exists some $g > 0$ for every $\varepsilon > 0$ so that,

$$\lim \sup_{n\to\infty} P(|E(U_\lambda^2) - E_{\tilde{n}}(U_\lambda^2)| > g\tilde{n}^{-1} \log n) < \varepsilon,$$

where $E_{\tilde{n}}(U_\lambda^2)$ is the empirical mean of U_λ in the sample of $\tilde{n}$ observations in the validation set. For the second and third term in the loss (23) it follows analogously that there exists $g > 0$ for every $\varepsilon > 0$ so that

$$\lim \sup_{n\to\infty} P(|E(U_\lambda V_\lambda) - E_{\tilde{n}}(U_\lambda V_\lambda)| > g\tilde{n}^{-1} \log n) < \varepsilon,$$

$$\lim_{n\to\infty} \sup P(|E(V_\lambda^2) - E_{\tilde{n}}(V_\lambda^2)| > g\tilde{n}^{-1}\log n) < \varepsilon.$$

Hence, using (23), there exists some $g > 0$ for every $\varepsilon > 0$ so that

$$\lim_{n\to\infty} \sup P\left(\sup_{\phi\in[0,1]} |L(\lambda,\phi) - L_{\tilde{n}}(\lambda,\phi)| > g\tilde{n}^{-1}\log n\right) < \varepsilon.$$

When extending the supremum over $\phi \in \Lambda$ to a supremum over $\lambda > 0,\ \phi \in [0, 1]$, note that it is sufficient, due to (22), to consider values of λ in the finite set $\{\lambda_1, \ldots, \lambda_m\}$. Using Bonferroni's inequality and $m \leqslant d \log n$, it follows that there exists some $g > 0$ for every $\varepsilon > 0$ so that

$$\lim_{n\to\infty} \sup P\left(\sup_{\lambda,\phi} |L(\lambda,\phi) - L_{\tilde{n}}(\lambda,\phi)| > g\tilde{n}^{-1}\log^2 n\right) < \varepsilon,$$

which completes the proof as $\tilde{n}/n \to 1/K > 0$ for $n \to \infty$. □

6.3. Proof of Theorem 5

For independent predictor variables, the loss $L(\lambda)$ of the Lasso estimator under penalty parameter λ is given by

$$L(\lambda) = \sum_{k\in\{1,\ldots,p\}} (\hat{\beta}_k^\lambda - \beta_k)^2 = \sum_{k\leqslant q} (\hat{\beta}_k^\lambda - \beta_k)^2 + \sum_{k>q} (\hat{\beta}_k^\lambda)^2, \tag{24}$$

using that the variance of all components of X is identical to 1 and $\beta_k = 0$ for all $k > q$. Let $\lambda_\star$ be defined as in (13). Using Lemma 14, it follows that for all $\varepsilon > 0$, with probability converging to 1 for $n \to \infty$, for all $k \leqslant q$ and $\lambda \geqslant \lambda_\star$,

$$(\hat{\beta}_k^\lambda - \beta_k)^2 \geqslant (1-\varepsilon)^2(\lambda - \lambda_\star)^2.$$

Summing only over components with $k \leqslant q$ in (24), it follows that the loss is bounded from below for $\lambda \geqslant \lambda_\star$ by

$$\inf_{\lambda\geqslant\lambda_\star} L(\lambda) \geqslant q(1-\varepsilon)^2\lambda_\star^2. \tag{25}$$

Now the case $\lambda < \lambda_\star$ is examined. The range of λ is furthermore restricted to lie in the area Λ, defined in (10). Denote in the following the difference between the Lasso estimators $\hat{\beta}^\lambda$ and $\hat{\beta}^{\lambda_\star}$ by $\delta^\lambda = \hat{\beta}^\lambda - \hat{\beta}^{\lambda_\star}$. Denote the difference between $\hat{\beta}^{\lambda_\star}$ and the true parameter β by $\theta = \hat{\beta}^{\lambda_\star} - \beta$. Then

$$(\hat{\beta}_k^\lambda - \beta_k)^2 = \theta_k^2 - 2\theta_k\delta_k^\lambda + (\delta_k^\lambda)^2.$$

It follows by Lemma 14 that, with probability converging to 1 for $n \to \infty$, for any $\varepsilon > 0$, $|\theta_k| > (1-\varepsilon)\lambda_\star$. It holds by an analogous argument that $|\theta_k| < (1+\varepsilon)\lambda_\star$. Hence, for all $k \leqslant q$,

$$(\hat{\beta}_k^\lambda - \beta_k)^2 \geqslant (1-\varepsilon)^2\lambda_\star^2 - 2(1+\varepsilon)\lambda_\star\delta_k^\lambda + (\delta_k^\lambda)^2.$$

By Lemma 11 and analogously to Lemma 14, it holds furthermore with probability converging to 1, that $(1-\varepsilon)(\lambda_0 - \lambda) \leqslant |\delta_k^\lambda| \leqslant (1+\varepsilon)(\lambda_0 - \lambda)$ and hence, for all $k \leqslant q$,

$$(\hat{\beta}_k^\lambda - \beta_k)^2 \geqslant (1-\varepsilon)^2\lambda_\star^2 - 2(1+\varepsilon)^2\lambda_\star(\lambda_\star - \lambda) + (1-\varepsilon)^2(\lambda_\star - \lambda)^2.$$

As $\lambda_\star$ is the largest value of λ such that $\mathcal{M}_\lambda = \mathcal{M}_\star$, a noise variable (with index $k > q$) enters the model $\mathcal{M}_\lambda$ if $\lambda < \lambda_\star$. Using again Lemma 10, with probability converging to 1 for $n \to \infty$, it holds for this component that for any $\varepsilon > 0$,

$$(\hat{\beta}_k^\lambda)^2 \geqslant (1-\varepsilon)^2(\lambda_\star - \lambda)^2. \tag{26}$$

It follows that with probability converging to 1 for $n \to \infty$,

$$L(\lambda) \geqslant q(1-\varepsilon)^2\lambda_\star^2 - 2q(1+\varepsilon)^2\lambda_\star(\lambda_\star - \lambda) + (q + (1-\varepsilon)^2)(\lambda_\star - \lambda)^2.$$

Denote the infimum over $\lambda_0 \leqslant \lambda \leqslant \lambda_\star$ of the r.h.s. by $f(\varepsilon)$,

$$f(\varepsilon) := \inf_{\lambda_0 \leqslant \lambda \leqslant \lambda_\star} (q(1-\varepsilon)^2\lambda_\star^2 - 2q(1+\varepsilon)^2\lambda_\star(\lambda_\star - \lambda) + (q + (1-\varepsilon)^2)(\lambda_\star - \lambda)^2).$$

Note that $f(\varepsilon)$ is a continuous function of ε and

$$\lim_{\varepsilon \to 1} f(\varepsilon) = q/(q+1)\lambda_\star^2.$$

Hence, as ε can be chosen arbitrarily close to 1, it holds that, with probability converging to 1 for $n \to \infty$,

$$\inf_{\lambda_0 \leqslant \lambda \leqslant \lambda_\star} L(\lambda) \geqslant \inf_{\varepsilon > 0} f(\varepsilon) \geqslant \lambda_\star^2/2.$$

By Lemma 13, $\lambda_\star^{-2} = \mathrm{O_p}(n^{1-\xi})$ and thus, using (25), for any $r > 1 - \xi$,

$$P\left(\inf_{\lambda \in \Lambda} L(\lambda) > cn^{-r}\right) \to 1, \quad n \to \infty,$$

which completes the proof. □

6.4. Proof of Theorem 7

It holds for the loss under $\hat{\lambda}$ and $\hat{\phi}$ for every $g > 0$ that

$$P(L(\hat{\lambda}, \hat{\phi}) > gn^{-1}\log^2 n) \leqslant P\left(\inf_{\lambda \in \Lambda, \phi \in [0,1]} L(\lambda, \phi) > \frac{1}{2} gn^{-1}\log^2 n\right) + 2P\left(\sup_{\lambda \in \Lambda, \phi \in [0,1]} |L(\lambda, \phi) - L_{\mathrm{cv}}(\lambda, \phi)| > \frac{1}{2} gn^{-1}\log^2 n\right).$$

It follows by Theorem 6 that the first term on the r.h.s. vanishes for $n \to \infty$. The second term is by Bonferroni's inequality bounded from above by

$$K \max_{1 \leqslant S \leqslant K} P\left(\sup_{\lambda \in \Lambda, \phi \in [0,1]} |L(\lambda, \phi) - L_{S,\tilde{n}}(\lambda, \phi)| > \frac{1}{2} gn^{-1}\log^2 n\right).$$

Using Lemma 15, there exists thus for every $\varepsilon > 0$ some $g > 0$ so that

$$\lim \sup_{n \to \infty} P(L(\hat{\lambda}, \hat{\phi}) > gn^{-1}\log^2 n) < \varepsilon,$$

which completes the proof. □

References

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–451.

Fan, J., Li, R., 2001. Variable selection via penalized likelihood. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.

Frank, I., Friedman, J., 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics 35, 109–148.

Huber, P., 1973. Robust regression: asymptotics, conjectures, and Monte Carlo. Ann. Statist. 1, 799–821.

Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Statist. 28, 1356–1378.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London.

Meinshausen, N., Bühlmann, P., 2006. High dimensional graphs and variable selection with the lasso. Ann. Statist. 34, 1436–1462.

Osborne, M., Presnell, B., Turlach, B., 2000. On the lasso and its dual. J. Comput. Graph. Statist. 9, 319–337.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58, 267–288.

Tsybakov, A., van de Geer, S., 2005. Square root penalty: adaptation to the margin in classification and in edge estimation. Ann. Statist. 33, 1203–1224.

van de Geer, S., van Houwelingen, H., 2004. High dimensional data: $p \gg n$ in mathematical statistics and bio-medical applications. Bernoulli 10, 939–943.

Zhao, P., Yu, B., 2004. Boosted lasso. Technical Report 678, University of California, Berkeley.