Electronic Journal of Statistics
Vol. 7 (2013) 1607–1631
ISSN: 1935-7524
DOI: 10.1214/13-EJS818

Sign-constrained least squares estimation for high-dimensional regression

Nicolai Meinshausen

Department of Statistics,
University of Oxford, UK
e-mail: meinshausen@stats.ox.ac.uk

Abstract: Many regularization schemes for high-dimensional regression have been put forward. Most require the choice of a tuning parameter, using model selection criteria or cross-validation. We show that a simple sign-constrained least squares estimation is a very simple and effective regularization technique for a certain class of high-dimensional regression problems. The sign constraint has to be derived via prior knowledge or an initial estimator. The success depends on conditions that are easy to check in practice. A sufficient condition for our results is that most variables with the same sign constraint are positively correlated. For a sparse optimal predictor, a non-asymptotic bound on the ℓ_1-error of the regression coefficients is then proven. Without using any further regularization, the regression vector can be estimated consistently as long as $s^2 \log(p)/n \to 0$ for $n \to \infty$, where s is the sparsity of the optimal regression vector, p the number of variables and n sample size. The bounds are almost as tight as similar bounds for the Lasso for strongly correlated design despite the fact that the method does not have a tuning parameter and does not require cross-validation. Network tomography is shown to be an application where the necessary conditions for success of sign-constrained least squares are naturally fulfilled and empirical results confirm the effectiveness of the sign constraint for sparse recovery if predictor variables are strongly correlated.

AMS 2000 subject classifications: Primary 62J07, 62J05.
Keywords and phrases: Variable selection, shrinkage estimators, quadratic programming, high-dimensional linear models.

Received November 2012.

Contents

1. Introduction

High-dimensional regression problems are characterized by a large number of predictor variables in relation to sample size. Regularization (in a broad sense) is of critical importance for high-dimensional problems and much attention has been paid to various schemes and their properties in recent years, including the *Ridge* estimator (Hoerl and Kennard, 1970), *non-negative Garrotte* (Breiman, 1995), the *Lasso* (Tibshirani, 1996) and various variations of the latter, including the *group Lasso* (Yuan and Lin, 2006), *adaptive Lasso* (Zou, 2006) and the more recent *square-root Lasso* (Belloni et al., 2011; Bunea et al., 2013). Datasets with very low signal-to-noise ratio offer similar challenges to high-dimensional problems even if the notional sample size is quite high.

Sign-constraints on the regression coefficients are a simpler regularization and have been first advocated by I.J. Good, as covered in the book Lawson and Hanson (1995). There is a wide range of problems where the sign of the regression coefficients can either be estimated by an initial estimator or where it is known a priori, such as in image processing and spectral analysis (Waterman, 1977; Bellavia et al., 2006; Donoho et al., 1992; Chen and Plemmons, 2009; Guo and Berman, 2012). Sign-constraints have also been implemented for matrix factorizations, specifically the *non-negative Matrix factorization* (Lee et al., 1999; Lee and Seung, 2001; Ding et al., 2010) and *non-negative least squares* regression can be a useful tool for this factorization (Kim and Park, 2007). We study the performance of non-negative least squares type problems under a so-called *Positive Eigenvalue Condition*, which can be checked for any given dataset by solving a quadratic programming problem. A sufficient condition uses only the minimum of all entries in the design matrix. It is shown that non-negative (or, in general, sign-constrained) least squares is a surprisingly effective regularization technique for high-dimensional regression problems under these conditions. If the *Positive Eigenvalue Condition* is not fulfilled, the sign constraint is still a good ingredient in a regularization framework. The *non-negative Garrote* (Breiman, 1995) is, for example, making use of a sign-constraint, where the signs are derived from an initial estimator as is the *positive Lasso* (Efron et al., 2004).

The data are assumed to be given by a $n \times 1$-vector of real-valued observations $\mathbf{Y}$ and a $n \times p$-dimensional matrix $\mathbf{X}$, where column k of $\mathbf{X}$ contains all n samples of the k-th predictor variable for $k = 1, \ldots, p$. The non-negative least squares

(NNLS) regression estimator is defined as

$$\hat{\beta} := \operatorname{argmin}_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \quad \text{such that} \quad \min_k \beta_k \geq 0. \tag{1}$$

We will work with a positivity constraint without limitation of generality since variables that are constrained to be negative can be replaced by their negative counterpart and the problem can thus always be framed as a non-negative least squares optimisation. Problem (1) is a convex optimization problem and can be solved with general quadratic programming problem solvers, including active set (Lawson and Hanson, 1995), iterative (Kim et al., 2006) and interior-point approaches (Bellavia et al., 2006). A tailor-made fast approximate algorithm based on random projections has recently been proposed in Boutsidis and Drineas (2009). The uniqueness of the solution has been studied in Bruckstein et al. (2008).

Note that the non-negative least squares estimator (1) does not require the choice of a tuning parameter beyond choosing the sign of the coefficients. Imposing a sign-constraint might seem like a very weak regularization but it will be shown that the estimator is remarkably different from the un-regularized least squares estimator. It can cope with high-dimensional problems where the number of predictor variables vastly exceeds sample size. It will be shown to be a consistent estimator as long as the underlying optimal predictor set is sufficiently sparse and the so-called *Positive Eigenvalue Condition* is fulfilled.

The recent manuscript Slawski et al. (2011) contains independent work on the behaviour of NNLS in high-dimensions, albeit from a slightly different viewpoint. Using the same *Positive Eigenvalue Condition* (which is called self-regularizing design condition), a bound on the prediction error of NNLS and a sparse recovery property after hard thresholding are shown in Slawski et al. (2011). While our main focus is on sparse recovery in the ℓ_1-sense (which is not covered in Slawski et al. (2011)) the bounds on prediction error are also of different nature since the assumptions are different. The compatibility condition is not used in the work (Slawski et al., 2011) and the prediction error bound is thus an order $\sqrt{n}$ larger than in our results. A more recent manuscript of the same authors (Slawski and Hein, 2012) studies the thresholded NNLS estimator in more detail, while also referring to the recovery results in the current manuscript. As with nearly all sparse recovery results in the ℓ_1-penalized estimation literature (Van De Geer and Bühlmann, 2009), our main result, the ℓ_1-bound in Theorem 1 depends critically on a compatibility condition, and we show that the same convergence rates as for Lasso estimation can be achieved with NNLS even though no penalty parameter has been used.

We stress that the achieved bounds on the ℓ_1-error can be achieved with Lasso-type estimators (e.g. Bickel et al., 2009) under similar if slightly weaker assumptions. The advantage of NNLS in practice is the lack of a tuning parameter. Cross-validation is not necessary for the NNLS estimator, whereas it would be hard to argue that one can dispense with it for practical Lasso applications. In this sense, it is interesting to see that NNLS can achieve similar bounds on the ℓ_1-error despite its simplicity.

The manuscript is organized as follows. The notation and the main two assumptions, the compatibility and *Positive Eigenvalue Condition*, are introduced in Section 2. Our main result, a ℓ_1-bound on the difference between the NNLS estimator and the optimal regression coefficients is shown in Section 3, along with a bound on the prediction error.

2. Notation and assumptions

We assume that the n samples $\mathbf{Y} \in \mathbb{R}^n$ are drawn from $\mathbf{X}\beta^* + \varepsilon$ for some p-dimensional vector β^* with $\min_k \beta_k^* \geq 0$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$. Let S be the set of non-zero entries of the optimal solutions, $S = \{k : \beta_k^* \neq 0\}$ and $N = S^c$ be the complement of S. We could also let β^* be the best approximation to the data-generating model under positivity constraints but will refrain from doing so for notational simplicity. We assume that the columns of $\mathbf{X}$ are standardized to ℓ_2-norm of n. Despite not necessarily assuming that the columns are mean-centered, we call $\hat{\mathbf{\Sigma}} = n^{-1}\mathbf{X}^T\mathbf{X}$ the covariance matrix throughout.

We make two major assumptions for the main result, a standard condition about sparse eigenvalues and another one about so-called positive eigenvalues.

2.1. Compatibility condition

There has been much recent work on the properties of the Lasso (Tibshirani, 1996). Many similar conditions for success of the Lasso penalization schemes have been derived (for example Zhang and Huang, 2008; Meinshausen and Yu, 2009; Wainwright, 2009; Bunea et al., 2007a,b; Van De Geer, 2008; Bickel et al., 2009). A good overview of all conditions and their relations is given in Van De Geer and Bühlmann (2009). The weakest condition is based on the notion of (L, S) restricted ℓ_1-eigenvalues. We first define the standard minimal eigenvalue of a matrix $\mathbf{A}$ as

$$\phi_{min}^2(\mathbf{A}) \;=\; \min\{\beta^T\mathbf{A}\beta : \|\beta\|_2^2 \leq 1\} \tag{2}$$

In a high-dimensional setting with $p > n$, this minimal eigenvalue will be 0 for the empirical covariance matrix $\hat{\mathbf{\Sigma}}$, although it will be bounded from below in general if we restrict $\mathbf{A}$ to small sub-matrices of the empirical covariance matrix.

The (L, S) restricted ℓ_1-eigenvalue of matrix $\mathbf{A}$ is defined as:

$$\phi_{compatible}^2(L, S, \mathbf{A}) := \min\left\{s\frac{\beta^T\mathbf{A}\beta}{\|\beta\|_1^2} : \beta \in \mathcal{R}(L, S)\right\}, \tag{3}$$

where $\mathcal{R}(L, S) = \{\beta : \|\beta_N\|_1 \leq L\|\beta_S\|_1\}$, $N = S^c$ and $s = |S|$.

A lower bound on this restricted eigenvalue is necessary for success of the Lasso, either in a prediction loss or coefficient recovery sense and was called the compatibility condition in Van De Geer and Bühlmann (2009). It was shown to be weaker than all similar conditions such as the *Restricted Isometry Property* (Candes and Tao, 2007).

We make the following assumption.

Assumption 1 (Compatibility Condition)**.** *There exists some* $\phi > 0$ *such that the* (L, S)*-restricted* ℓ_1*-eigenvalue* $\phi^2_{compatible}(L, S, \hat{\mathbf{\Sigma}}) \geq \phi$.

The value of L will be specified in Theorem 1.

The assumption is formulated for the empirical covariance matrix $\hat{\mathbf{\Sigma}}$ but can also easily be reformulated on the population covariance matrix $\mathbf{\Sigma}$ for random design. Assume that the maximal difference between the population and empirical covariance matrix is bounded by $\delta > 0$, that is $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq \delta$. This assumption is fulfilled with high probability for many data sets with larger sample size. If the predictors have for example a multivariate normal distribution (which will not be assumed elsewhere), then the condition is fulfilled with probability $1 - 2\exp(-t)$ for $\delta \geq \sqrt{u} + u$ with $u = (4t + 8\log(p))/n$, see (10.1) in Van De Geer and Bühlmann (2009). If $\delta \leq \phi^2/(4(L+1)^2 s)$, then $\phi^2_{compatible}(L, S, \mathbf{\Sigma}) \geq \phi$ implies $\phi^2_{compatible}(L, S, \hat{\mathbf{\Sigma}}) \geq \phi/2$. The proof follows from the inequality $\phi^2_{compatible}(L, S, \hat{\mathbf{\Sigma}}) \geq \phi^2_{compatible}(L, S, \mathbf{\Sigma}) - (L+1)\sqrt{\delta s}$ in Corrolary 10.1 in Van De Geer and Bühlmann (2009). The *Compatibility Condition* could thus be imposed on the population covariance matrix instead of the empirical covariance matrix.

2.2. Positive eigenvalue condition

The following *Positive Eigenvalue Condition* is the central assumption for the main result.

The positively constrained minimal ℓ_1- eigenvalue of matrix $\mathbf{A}$ is defined as

$$\phi^2_{pos}(\mathbf{A}) := \min\left\{\frac{\beta^T \mathbf{A}\beta}{\|\beta\|_1^2} : \min_k \beta_k \geq 0\right\}, \tag{4}$$

A lower bound on this restricted eigenvalue will be a sufficient condition for sparse recovery success of NNLS.

Assumption 2 (Positive Eigenvalue Condition)**.** *There exists some* $\nu > 0$ *such that* $\phi^2_{pos}(\hat{\mathbf{\Sigma}}) \geq \nu$.

A lower bound on this eigenvalue seems to be a strong condition. However, the *Positive Eigenvalue Condition* is restricted to positive coefficients. There are thus some immediate examples where it is fulfilled, which we discuss below.

Example I: strictly positive covariance matrix. The *Positive Eigenvalue Condition* is fulfilled if $\min_{i,j} \hat{\mathbf{\Sigma}}_{i,j} \geq \nu > 0$, that is all entries in the covariance matrix are strictly positive. Again, this condition could also be formulated for the population covariance matrix, using a bound on $\|\mathbf{\Sigma} - \hat{\mathbf{\Sigma}}\|_\infty$.

We also remark on the case of general sign-constraints (some variables constrained to be positive, some negative). The condition applies then to the dataset where all variables with a negativity constraint have been replaced with their negative counterparts. The constraint on the original covariance matrix is thus that it forms two blocks. The variables in the first block are the variables with

a positivity constraint and the second block is formed by all variables with a negativity constraint. Correlations are required to be positive within a block and negative between blocks.

A generalization of Example I is the following.

Example II: only few negative entries. Let $\mathcal{A} := \{i : \hat{\mathbf{\Sigma}}_{ij} < 0 \text{ for some } 1 \leq j \leq p\}$ be the minimal set such that $\hat{\mathbf{\Sigma}}_{ij} < 0$ implies $\{i, j\} \subseteq \mathcal{A}$ for all $1 \leq i, j \leq p$. The *Positive Eigenvalue Condition* is fulfilled if both of the conditions below are fulfilled for some $\nu > 0$.

1. All entries of the covariance matrix are strictly positive on $\mathcal{A}^c$, that is $\hat{\mathbf{\Sigma}}_{ij} \geq 2\nu$ if $\{i, j\} \subseteq \mathcal{A}^c$ for all $1 \leq i, j \leq n$.
2. A restricted eigenvalue condition holds on the set $\mathcal{A}$, ie

$$\min\Big\{\frac{\beta^T \hat{\mathbf{\Sigma}}\beta}{\|\beta\|_1^2} : \min_k \beta_k \geq 0 \text{ and } \beta_k = 0 \text{ for all } k \in \mathcal{A}^c\Big\} > 2\nu.$$

If the set $\mathcal{A}$ is very small, in particular much smaller than n, the latter restricted ℓ_1-eigenvalue condition is in general not very restrictive. The important criterion is thus whether the set $\mathcal{A}$ is small compared to the sample size.

Example III: block matrix. For a $p \times p$-matrix $\mathbf{A}$ and a set $K \subseteq \{1, \ldots, p\}$, let $\mathbf{A}_{KK}$ be the $|K| \times |K|$-submatrix formed by all elements in set K. Suppose

1. Entries of the covariance matrix can be negative but fulfil $\hat{\mathbf{\Sigma}}_{ij} \geq -\rho/p^2$ for all $1 \leq i, j \leq n$ and some $\rho > 0$.
2. The set of variables $\{1, \ldots, p\}$ can be partitioned into $B \geq 1$ blocks $B_j \subseteq \{1, \ldots, p\}$ such that $\phi^2_{pos}(\hat{\mathbf{\Sigma}}_{B_j B_j}) \geq (\nu + \rho)B$ for all $j = 1, \ldots, B$.

A more specific example is thus: all entries in $\hat{\mathbf{\Sigma}}$ are larger than $-\rho/p^2$ for some $\rho > 0$ and $\hat{\mathbf{\Sigma}}_{ij} \geq (\nu + \rho)B$ if both i, j are within the same block.

The positive aspect of the condition is that it is very easy to check in practice whether it applies (at least approximately) and whether one would thus expect the bounds shown below to apply to a given dataset.

3. Main results

It will be shown that non-negative least squares leads to a good recovery of the optimal sparse regression vector for high-dimensional data. We study the ℓ_1-error in the regression vector, which also yields a bound on the ℓ_2-error and prediction loss.

Theorem 1. *Assume that the* Positive Eigenvalue Condition *holds with $\nu > 0$. Choose any $0 < \eta < 1/5$. Assume that the compatibility condition holds with $\phi_\nu > 0$ for $L_\nu = 3/\sqrt{\nu}$ and the compatibility condition holds with $\phi_\infty \geq \phi_\nu$ for $L_\infty = 0$. Set*

$$K^2 := 2\log\big(\frac{p}{\sqrt{2\pi}\eta}\big).$$

It then holds with probability at least $1-\eta$ *that*

$$\|\hat{\beta}-\beta^*\|_1 \;\leq\; \frac{2Ks\sigma}{\sqrt{\phi_\infty}n}\Big(1+\max\{\frac{3s}{\phi_\nu},\frac{12}{\nu}\}\Big).$$

If, additionally, the minimal eigenvalue $\phi^2_{min}(\hat{\mathbf{\Sigma}}_{SS})$ *is greater than or equal to some* $\tau>0$ *and* $\min_{k\in S}\beta_k > K\sigma/\sqrt{n\tau}$*, with probability at least* $1-\eta$*,*

$$\|\hat{\beta}-\beta^*\|_1 \;\leq\; \frac{2Ks\sigma}{\sqrt{\phi_\infty}n}\Big(1+\max\{\frac{\sqrt{\phi_\infty}}{\phi_\nu},\frac{4\sqrt{\phi_\infty}}{s\nu}\}\Big).$$

A proof is given in the appendix.

The result might be surprising since it implies (for sufficiently large positive coefficients) that non-negative least squares is succeeding in recovering the regression coefficients in an ℓ_1-sense if $s^2\log(p)/n\to 0$ for $n\to\infty$ if keeping all other variables constant, a scaling that requires for general design a lot more regularization in the form of Lasso penalties (or similar). Some comments:

1. The NNLS estimator can thus attain good estimation accuracy for high-dimensional data in the absence of a strong regularization like ℓ_1-constraints on the coefficient vector. An intuitive explanation is that the positivity constraint acts comparable to a ℓ_1-constraint if the minimal positive eigenvalue is bounded away from 0. To take a simple example, use the previous Example III of a block matrix where all entries in $\hat{\mathbf{\Sigma}}$ are greater than $\rho\in(0,1)$. Then $\|\mathbf{X}\beta\|_2^2$ is for every non-negative vector β at least $\rho\|\beta\|_1^2$. A bound on the ℓ_2-norm of the prediction $\|\mathbf{X}\beta\|_2^2$ acts thus also as a constraint on the ℓ_1-norm of β_1 and might help to illustrate (for this particular example) why NNLS can achieve similar estimation accuracy as Lasso estimation.
2. The assumption made for the second part of the theorem bounds the smallest positive regression coefficient away from 0. This assumption leads to an equivalence of the oracle and restricted least-squares estimator and allows tighter bounds but consistent estimation is also possible in the absence of this assumption, as shown by the first part of Theorem 1.
3. The constant $L_\nu=3/\sqrt{\nu}$ in the required compatibility condition can get large for small values of the minimal positive eigenvalue ν. In general, NNLS will cease to be an interesting procedure for $\nu\to 0$, as also shown in the section with numerical results. As an extreme case, imagine for every column in $\mathbf{X}$ there exists also the negative of it as a column, which puts the positive eigenvalue ν to 0. NNLS is identical to OLS estimation in this setting and will in general not succeed in a high-dimensional setting. On the other hand, as long as ν is bounded away from 0 (as in Examples I-III), the constant $L_\nu=3/\sqrt{\nu}$ can be of a reasonable size but will even then be larger than the standard value of 3 in the case of ℓ_1-constrained estimation.
4. A similar result will be shown in Section 3.3, where the assumption on the minimal positive eigenvalue will be replaced by a slightly strength-

ened assumption. Under the strengthened assumption, the compatibility condition only needs to hold for $L = 0$ instead of $L_\nu = 3/\sqrt{\nu}$.

We will first show two implications of Theorem 1 on sign recovery and prediction error.

3.1. Sign recovery

The result does not imply exact sign recovery in the sense that the non-zero coefficients equal exactly the set S (and indeed this will in general not be the case), but it implies that the s largest coefficients correspond to the variables in the set S.

Corollary 1. *If all conditions in Theorem 1 are fulfilled and the stronger assumption that the minimum over all non-zero coefficients is bounded from below by the maximum of* $K\sigma/\sqrt{n\tau}$ *and*

$$\min_{k\in S} \beta_k \geq \frac{2Ks\sigma}{\sqrt{\phi_\infty n}}\Big(1 + \max\{\frac{\sqrt{\phi_\infty}}{\phi_\nu}, \frac{4\sqrt{\phi_\infty}}{s\nu}\}\Big),$$

it holds with probability at least $1 - \eta$ *that the indices of the* s *largest absolute coefficients in* $\hat{\beta}$ *are identical to the set* S.

This follows immediately from Theorem 1 since the ℓ_1-bound on the difference between $\hat{\beta}$ and β^* implies the same bound in the supremum-norm.

3.2. Prediction error

The bound in Theorem 1 also implies a bound on the prediction error. Let $N = \{1, \ldots, p\} \setminus S$ be the set of noise variables with exactly vanishing coefficient $\beta^*_N \equiv 0$ and define the oracle estimator $\hat{\beta}^{oracle}$ for a set of noise variables N as

$$\hat{\beta}^{oracle} := \operatorname{argmin}_\beta \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \quad \text{such that} \quad \min_k \beta_k \geq 0 \ \text{ and } \ \beta_N \equiv 0. \tag{5}$$

The prediction error can then be bounded as follows.

Theorem 2. *If all conditions in Theorem 1 are fulfilled, with probability at least* $1 - \eta$ *for any* $0 < \eta < 1/5$,

$$\|\mathbf{X}(\hat{\beta}^{oracle} - \hat{\beta})\|_2^2 \ \leq \ \frac{2K^2\sigma^2}{n} \max\{\frac{2s}{\phi_\nu}, \frac{8}{\nu}\}.$$

A proof is given in the appendix. The mean squared error, introduced by using NNLS instead of the oracle estimator is thus bounded by $\log(p)s/n$ under the stronger assumption on the minimal non-zero regression value.

It is interesting to note what happens under model misspecification, that is if some variables are assigned a sign constraint that is opposite to the sign of the optimal regression vector. Let $\tilde{\beta}$ be the optimal regression vector in a population

sense such that all imposed sign constraints are satisfied for $\tilde{\beta}$. Note that the predictive accuracy of $\tilde{\beta}$ is at least as good as when using β', which is defined to be equal to β^* for all correctly specified variables and equal to 0 for all variables with a misspecified sign. If we let the oracle estimator be defined as in (5) but with the set of noise variables N now defined as the set of zero coefficients of $\tilde{\beta}$ instead of the set of zero coefficients of β^*, then we will find the same result as in Theorem 2 for this misspecified vector. If a set of variables have thus misspecified sign, we effectively lose the contribution of the misspecified variables and the predictive accuracy will then be identical to the case where the misspecified variables would not have appeared in the original dataset.

3.3. Alternative assumption

We can replace the compatibility condition (3) with a potentially larger value of L and the assumption on the mimimal positive eigenvalue (4) by an assumption on the minimal positive compatible eigenvalue, defined as

$$\phi^2_{pos,S}(\mathbf{A}) := \min\Big\{\frac{\beta^T\mathbf{A}\beta}{\|\beta\|_1^2} : \min_{k\in N}\beta_k \geq 0\Big\}. \tag{6}$$

In contrast to the minimal positive eigenvalue (4), the coefficients of β are only constrained to be positive outside of $S := \{k : \beta_k \neq 0\}$. The assumption of a bound on (6) is thus stronger than a bound on the minimal positive eigenvalue (4) but allows to assume $L = 0$ in the compatibility condition.

Theorem 3. *Assume the compatibility condition holds for $L_\infty = 0$ for some $\phi_\infty > 0$, that is $\phi^2_{compatible}(0, S, \hat{\mathbf{\Sigma}}) \geq \phi_\infty$. Assume that $\phi^2_{pos,S}(\hat{\mathbf{\Sigma}}) \geq \kappa$ for some $\kappa > 0$. Set K^2 again equal to $2\log\{p/(\sqrt{2\pi}\eta)\}$ for some $0 < \eta < 1/5$. With probability at least $1 - \eta$,*

$$\|\hat{\beta} - \beta^*\|_1 \;\leq\; \frac{8Ks\sigma}{\kappa\sqrt{\phi_\infty n}}.$$

A proof is given in the appendix.

A lower bound on the minimal positive compatible eigenvalue seems reasonable under some scenarios. To give the simplest, consider the setting where there exists a set $A \subseteq \{1, \ldots, p\}$ such that $S \subseteq A$ and (a) the minimal positive eigenvalue (4) is greater or equal than some $\nu > 0$, (b) the minimal eigenvalue $\phi^2_{\min}(\hat{\mathbf{\Sigma}}_{AA})$ is greater or equal to some $\kappa > 0$ and (c) all variables in A are orthogonal to variables in A^c. The positive compatible eigenvalue (6) is then at least $\min\{\mu, \kappa\}$. Theorem 3 thus shows that the large constant $L_\nu > L_\infty$ in the compatibility condition is not necessary in all scenarios. In general, however, it will be more difficult to verify a bound on the minimal positive compatible eigenvalue (6) than on the minimal positive eigenvalue (4) as used in Theorem 1. We will thus mainly work with (4) and Theorem 1 which states that NNLS will be competitive with ℓ_1-constrained estimation if the value ν in the original minimal positive eigenvalue assumption (4) is not very small. If ν is

indeed well bounded away from 0 (but it will always be less than 1 by the normalization), the required constant $L_\nu = 3/\sqrt{\nu}$ in the compatibility condition will then just be a small factor larger than for the standard Lasso estimator and the stronger structural assumption of a bound on (6) will not be necessary. The following section with numerical results will illuminate some of these issues from an empirical point of view.

4. Numerical results

4.1. Toeplitz design

The theoretical results indicate that NNLS will be more competitive with ℓ_1-constrained estimation if the correlation between variables is very strong, as this will raise the *minimal positive eigenvalue* ϕ_{pos} as defined in (4) and in turn also weaken the required compatibility condition. Looking at the simplest possible example, we use a Toeplitz design for the population covariance matrix by setting $\mathbf{\Sigma}_{kk'} = \rho^{|k-k'|/p}$ for $k, k' \in \{1, \ldots, p\}$ and some value $\rho \geq 0$. For $\rho = 0$, all variables are sampled independently and we would expect NNLS to be less competitive with ℓ_1-constrained estimation than for large values ρ. We simulate $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where β is identically zero, except for $s \geq 1$ randomly chosen coefficients that are all set to 1 for $s \in \{3, 10, 20\}$. We also add two simulations where the optimal regression vector is not sparse in an ℓ_0-sense. In the first, the coefficient of β^* are distributed iid $\exp(\mu)$ for some value of μ (the results will be independent of the scaling μ and we chose $\mu = 1$). In the second we sample the coefficients β_k^* as the absolute value of independent samples from a Cauchy-distribution. The standard deviation of the normally distributed noise is set to σ^2 times the standard deviation of the signal contribution $\mathbf{X}\beta$.

Results are shown in Figure 1. We plot the ratio of the ℓ_1-approximation error between the NNLS estimator and Lasso estimator with a cross-validated choice of the penalty parameter λ. There are two main observations:

(a) The relative estimation error of the NNLS estimator (when compared to the cross-validated Lasso estimator) is smaller when the signal-to-noise ratio is high (low values of σ). This is to be expected as the coefficients of NNLS-selected variables are estimated with least squares estimation. The additional shrinkage Lasso applies to selected coefficients can be detrimental if the signal-to-noise ratio is high and a re-estimation of the selected model would be preferable. Many such two-stage procedure exist for Lasso-type estimation (Zou, 2006; Meinshausen, 2007; Candes and Tao, 2007). No such re-estimation is necessary for the Lasso for high signal-to-noise ratio data. On the other hand, the NNLS results can suffer from a higher variance for low signal-to-noise ratio data (high σ).

(b) For low correlation ρ between variables in the Toeplitz design, Lasso typically outperforms NNLS estimation. This is expected from the theory as the conditions get more stringent if the minimal positive eigenvalue is getting very close to 0. The minimal positive eigenvalue for the population

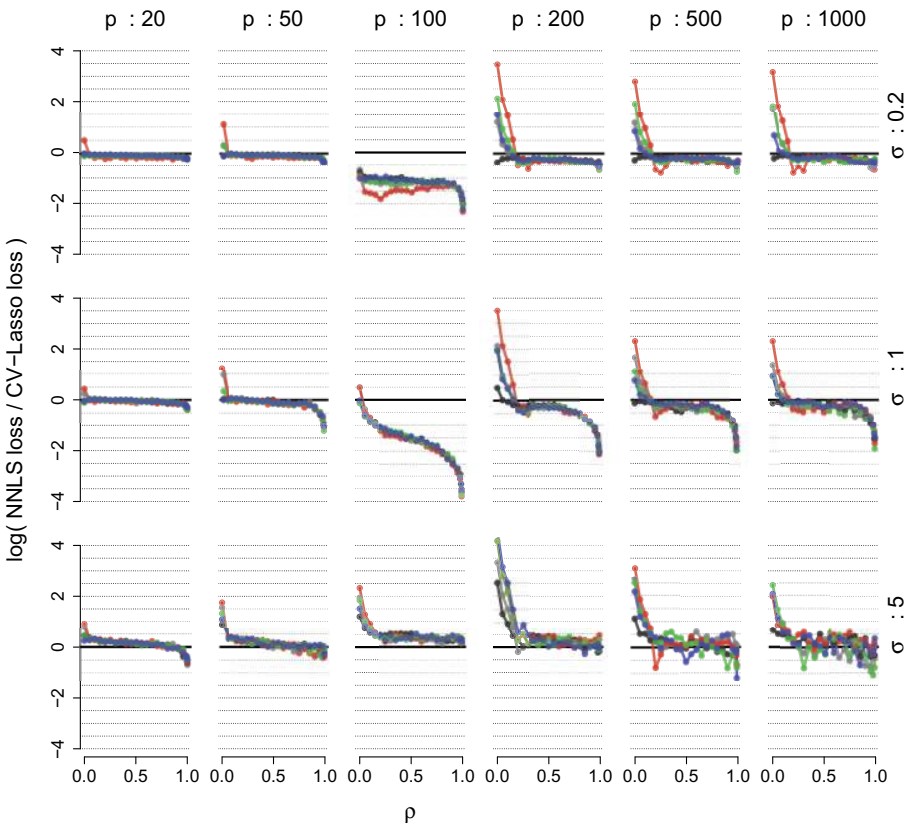

FIG 1. *For various values of p and σ, the logarithm of the ratio $\|\hat{\beta} - \beta^*\|_1 / \|\hat{\beta}^\lambda - \beta^*\|_1$ is shown as a function of $\rho \in [0, 1)$ in a Toeplitz design, where $\hat{\beta}$ is the NNLS estimator and $\hat{\beta}^\lambda$ a Lasso-estimator with a cross-validated choice of λ. The five lines correspond to sparsities $s = 3$ (red), $s = 10$ (green), $s = 20$ (blue), exponentially- (light gray) and Cauchy- (dark grey) distributed regression coefficients. Values above 0 indicate that the ℓ_1-loss of the Lasso estimation is smaller than the corresponding loss of the NNLS estimator and vice versa. A low signal-to-noise ratio (high σ) favours the Lasso and a high signal-to-noise ratio (low σ) favours NNLS, as the selected coefficients are identical to least squares estimates with the latter while additional shrinkage is applied when using the Lasso. The NNLS estimator is very competitive for heavily correlated design (large values of ρ) and suffers for smaller values of nearest-neighbour correlation ρ, as expected from the theory where stronger assumptions are necessary for the NNLS estimation error for smaller values of ρ.*

covariance matrix is bounded for Toeplitz design from below by ρ. On the other hand, for large values of ρ, the minimal positive eigenvalue becomes bounded well away from 0 and NNLS typically performs similarly as cross-validated Lasso estimation, even for rather low signal-to-noise ratio settings despite its simplicity. For values of p around the sample size, NNLS can even substantially outperform cross-validated Lasso estimation.

We look next at block design and show thereafter the network tomography example.

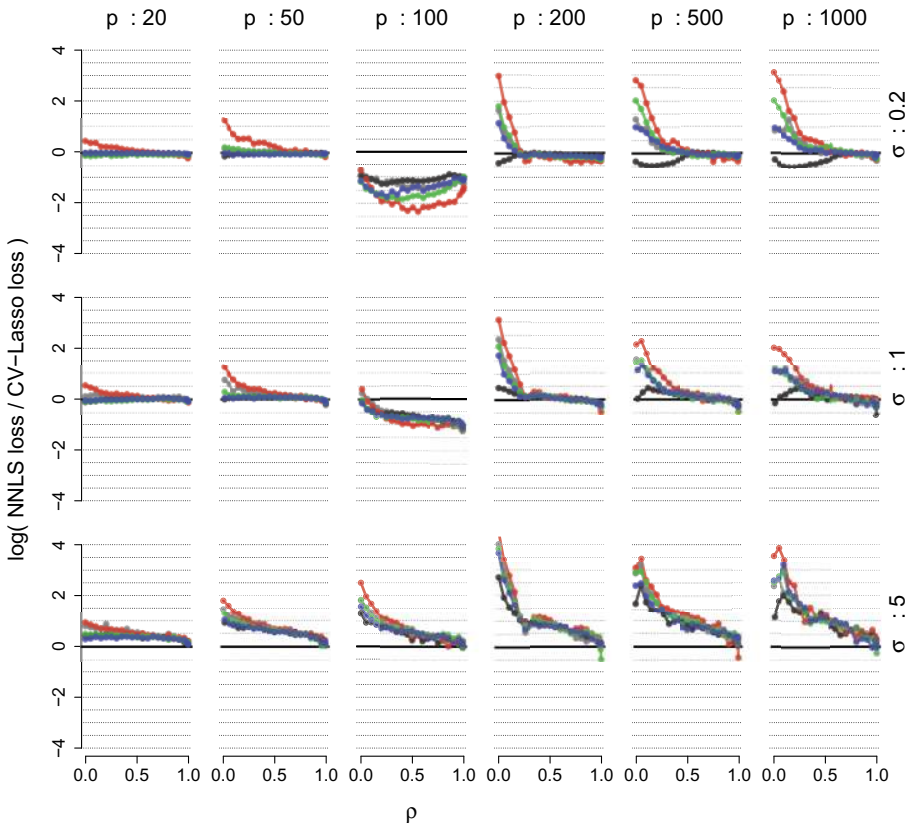

FIG 2. *Analogous plot to Figure 1 for block design with five independent blocks of predictor variables. For various values of p and σ, the logarithm of the ratio* $\|\hat{\beta} - \beta^*\|_1 / \|\hat{\beta}^\lambda - \beta^*\|_1$ *is shown as a function of ρ, where* $\rho \in [0, 1)$ *is the correlation between all predictors in the same block. The NNLS estimator is again very competitive for high correlation (large values of ρ), where the* ℓ_1*-approximation error is very close to the Lasso error, corresponding to values around 0 in the graph.*

4.2. Block design

We repeat the simulation of the previous section for a block design. The population covariance matrix is set to 0, except for five blocks of equal size along the diagonal, where off-diagonal elements are set to $\rho \in [0, 1)$. Diagonal elements of the population covariance matrix are set to 1. The regression vector and noise is chosen as in the previous section. This design is a specific case of the previous Example III and the positive eigenvalue condition holds true for $\nu = \rho/5$.

Results are shown in Figure 2. As for Toeplitz design, a high correlation between variables makes the NNLS estimator competitive with a cross-validated Lasso estimator. The relative error can be smaller than Lasso estimation for moderate to low-signal-to-noise ratios or very high values of the correlation. For $\rho = 0$, all variables are independent and NNLS in generally has a ℓ_1-error that is inflated by a potentially large factor when compared to cross-validated Lasso

estimation. For $\rho \geq 0.5$ and $\sigma \leq 1$, NNLS has typically a very similar ℓ_1-loss compared with Lasso estimation. For values of p around the sample size $n = 100$, NNLS can have a much better accuracy. The implication is again that NNLS is a suitable procedure for high-dimensional estimation for strongly correlated design with moderate to high signal-to-noise ratios.

4.3. Network tomography

The results above imply that NNLS can be very effective if (a) the sign of regression coefficients is known or can easily be estimated and (b) the *Positive Eigenvalue Condition* holds. *Network tomography* is a good example. For others, including image analysis and applications in signal processing, see Slawski et al. (2011). There are different aspects of network tomography, including origin-destination matrix estimation and link-level network tomography; see Castro et al. (2004) for a good overview of the statistical aspects and Xi et al. (2006) and Lawrence et al. (2006) for a discussion of active tomography in the context of link-level analysis. We will focus on the aspect of the link-level network tomography. The network consists of nodes arranged in a directed acyclic graph (or sometimes as a special case a tree) and measurements can be taken at the leaf nodes. These measurements are used to infer the state of all the nodes in the network. In a communication network, the measurements can be the delay or loss rate of packages. In a transport network (such as water distributions networks), it can be the shortfall of the flow rate compared to the expected rate. Since the network topology is assumed to be known, the measurements consist typically of noise plus a linear combination of the internal and unobservable states of the nodes in the network. If a node in the network has a loss (be it in the form of delaying packages or loss of water flow), it will have a linear effect on all leaf nodes that are descendants of the node in the directed acyclic graph.

Figure 3 shows a toy example. Imagining a flow passing through the tree from the internal nodes to the leaf nodes, the entry $\mathbf{X}_{i,j}$ is the proportion of flow in node j that reaches leaf node i if flow is divided equally among all outgoing edges in each node of the tree. Three internal nodes have loss rates $(\beta_1, \beta_2, \beta_3) = (10, 10, 0)$. The loss rates $\mathbf{Y} = (Y_1, Y_2, Y_3)$ at the three leaf nodes are then given by $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ for some i.i.d. noise ε and

$$\mathbf{X} = \begin{pmatrix} 0.3 & 0.5 & 0 \\ 0.3 & 0 & 0.5 \\ 0.4 & 0.5 & 0.5 \end{pmatrix}.$$

A positivity constraint on the coefficient vectors is clearly appropriate since there will in general not be a negative loss at internal nodes (for example no unexpected *gain* of water in a distribution network). In the noiseless case, the NNLS solution recovers exactly the internal states $(10, 10, 0)$ and thus identifies correctly the first two nodes as responsible for the loss of the flow rate in all three leaf nodes. In this simple example, the number of leaf nodes is equal to the number of internal nodes and ordinary least squares would also work in

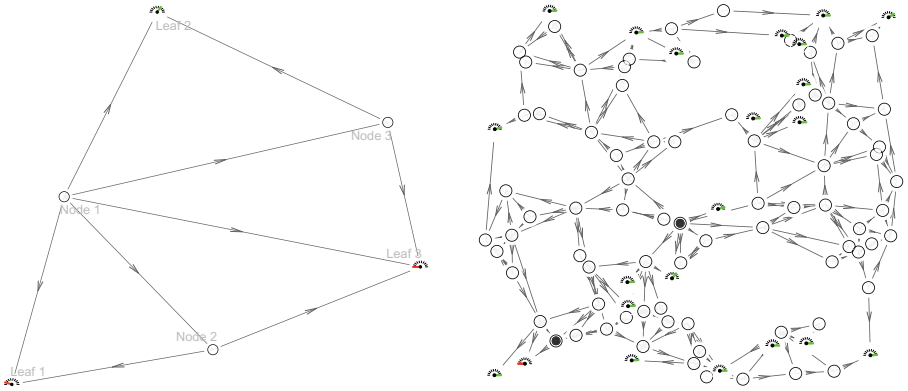

FIG 3. *Left: A network with three internal nodes and three leaf nodes. The (unobservable) losses at the internal nodes are (10,10,0), meaning that the first two nodes lead to a loss rate of 10 and the third node is not leading to any losses. The observations of the loss rates at the leaf nodes are then (8,3,9). Using the observations at the leaf nodes and knowledge of the topology, NNLS can correctly identify the two first nodes as responsible for the losses. Right: A network with 78 internal nodes and 22 leaf nodes. Two of the internal nodes have a positive loss (marked with a dot) and the observations at the leaf nodes are again sufficient to pinpoint the (unknown) location of the two nodes using NNLS estimation.*

the noiseless case. Least squares clearly ceases to be useful once the number of internal nodes exceeds the number of leaf nodes. Note that, contrary to the previous literature (for example Castro et al. (2004); Lawrence et al. (2006)) we do not attempt to fit a stochastic model to the observations. We are merely trying to directly estimate the current internal state β of the nodes in the network as accurately as possible.

The theory suggests that a non-negativity constraint can already be very powerful under certain constraints on the design matrix. The main condition is the *Positive Eigenvalue Condition*. In our simple network tomography example, it is obvious that all entries in $\mathbf{X}$ are non-negative and the same is hence true for $\hat{\mathbf{\Sigma}} = n^{-1}\mathbf{X}^T\mathbf{X}$. Entries in $\mathbf{X}$ correspond to the amount of loss (delay of packages or reduction in flow rate) in a leaf node caused by a specific loss at an internal node and is non-zero if and only if there is a connection between the internal and the leaf node. Suppose that all non-zero entries in $\mathbf{X}$ have entries at least as large as δ for some $\delta > 0$. Suppose further that we can group all internal nodes into B blocks such that the internal nodes within a block share at least one leaf node to which they all connect. The *Positive Eigenvalue Condition* is then fulfilled with value δ^2/B; see Example III in the discussion of the condition.

The theory seems to show that under these conditions the NNLS-regularization is effective. To test this, we examine the effect of placing an additional ℓ_1-constraint on the coefficient by computing

$$\beta^\lambda := \operatorname{argmin}_\beta \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \quad \text{such that } \min_k \beta_k \geq 0 \text{ and } \|\beta\|_1 \leq \lambda. \tag{7}$$

Let $\hat{\beta}$ be again the NNLS-solution defined in (1). It is obvious that $\beta^\lambda \equiv \hat{\beta}$ for all $\lambda \geq \lambda_{\max}$ for $\lambda_{\max} := \|\hat{\beta}\|_1$.

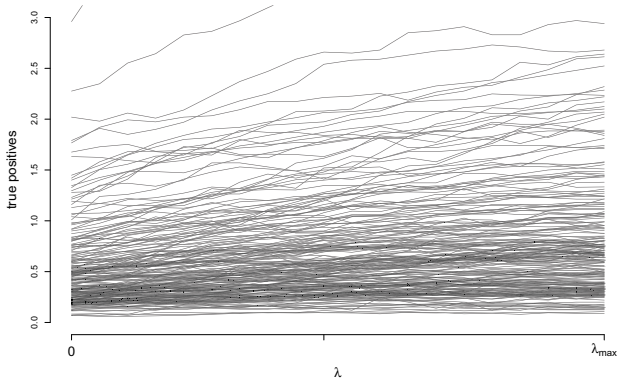

FIG 4. *The average number of correctly identified internal nodes with a positive loss under 1000 different scenarios with an additional ℓ_1-constraint as in (7). The NNLS solution corresponds to $\lambda = \lambda_{\max}$ and is seen to be in general superior to the solutions under additional shrinkage.*

We generate networks of similar type as the ones shown in Figure 3. The number N of total nodes is chosen for each of 1000 simulations uniformly out the set $\{25, 50, 100, 200, 400\}$. Nodes are distributed uniformly on the area $[-1, 1]^2$ and numbered in order of their Euclidean distance from the origin. Starting with the first node $k = 1$ closest to the origin, edges are drawn between it and its K nearest neighbours with a larger ordering number (where K is drawn uniformly from the set $\{5, 10, 20\}$). When drawing edges at node $k = 1, \ldots, N - 1$, they are deleted with probability κ (where κ is drawn uniformly from the set $\{.2, .4, .6, .8, 1\}$) or when the edge would cross a previously drawn edge. Imagining again a flow passing through the tree from the internal nodes to the leaf nodes, the entry $\mathbf{X}_{i,j}$ is the proportion of flow in node j that reaches leaf node i if flow is divided equally among all outgoing edges in each node of the tree. For each of the 1000 simulations, we draw a single graph from the parameters as described above and also draw the noise variance uniformly from the set $\{0, 0.125, 0.25, 0.5, 1, 2, 4\}$ and a number s of non-zero entries in β (corresponding to nodes with a delay or loss), where s is drawn uniformly from the set $\{2, 5, 10\}$. The s non-zero entries from β are generated independently as the absolute value of a standard-normal random variable. For each such setting, we simulate 50 times the vector $\mathbf{Y}$ and reconstruct with $\hat{\beta}^\lambda$ as in (7) for an evenly spaced grid of 20 points between $\lambda = 0$ and $\lambda = \lambda_{\max}$, the NNLS solution. Nodes are put in decreasing order of the reconstructed value $\hat{\beta}^\lambda$. We record the first entry in the re-ordered vector $\hat{\beta}^\lambda$ that corresponds to a false positive (a zero entry in the equally re-ordered vector β) and call the number of true positives the number of values of $\hat{\beta}^\lambda$ with larger value than the first false positive.

Figure 4 shows the average number of true positives as a function of λ. Each line corresponds to the average value over all 50 simulations in a given scenario.

For nearly all scenarios there is no benefit in placing an additional ℓ_1-penalty on the coefficients. The NNLS solution is thus a very good and simple estimator in these settings, as expected from theory. Additional regularization by an ℓ_1-penalty does not seem to improve results.

5. Discussion

We have shown that non-negative (or sign-constrained) least squares can be an effective regularization technique for sparse high-dimensional data under two conditions: (a) the data fulfil the so-called *Positive Eigenvalue Condition*, which is easy to check for a given dataset and (b) the sign of the coefficients is known or can easily be estimated.

If the conditions hold, NNLS can recover the correct sparsity pattern in the absence of any further regularization, as long as $s^2\log(p)/n \to 0$ for $n \to \infty$, where p is the number of variables, s the number of non-zero variables in the optimal regression vector and n is sample size. The standard *Compatibility Condition* is required for the results with a potentially large value of the constant. We have shown that the assumption just has to hold with a value $L = 0$ if the *Positive Eigenvalue Condition* is strengthened appropriately.

We have shown network tomography as an example where the sign of regression coefficients is known a priori and the design condition is fulfilled automatically, at least approximately. In other examples the sign can be estimated by an initial estimator. Estimation of the signs can be based for example on a tuning-parameter free method such as marginal regression or a Basis Pursuit. An attractive feature of NNLS is that it does not require any tuning parameter beyond the choice of the signs of the individual regression coefficients. Despite its simplicity, it can be remarkably accurate for high-dimensional regression, especially for strongly correlated design and moderate to high signal-to-noise ratios.

6. Appendix: Proofs

6.1. Proof of Theorem 1

First, for any $C > 0$, $1 - \Phi(C) \leq (2\pi)^{-1/2}C^{-1}\exp(-C^2/2)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Choosing $C^2 = K^2 = 2\log(\frac{p}{\sqrt{2\pi}\eta})$, it follows with $\eta < 1/5$ and hence $C \geq 1$ that $1 - \Phi(C) \leq \eta/p$. Thus $1 - p(1 - \Phi(C)) \geq 1 - \eta$ and the results follow hence from Lemma 1.

6.2. Proof of Theorem 2

Define the oracle non-negative least squares as in (5) and let $\delta\beta = \hat{\beta} - \hat{\beta}^{oracle}$. Let M be the set $M := \{k : \delta\beta_k < 0\}$. Using Equation (15) in the proof of

Lemma 1, it follows that, with probability at least $1-p(1-\Phi(C))$ (where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution),

$$\|\mathbf{X}(\hat{\beta}^{oracle}-\hat{\beta})\|_2^2 = \delta\beta^T\hat{\mathbf{\Sigma}}\delta\beta \;\leq\; 2\frac{C\sigma\gamma}{\sqrt{n}}\|\delta\beta_{M^c}\|_1,$$

where $\gamma = 1$ if the bound on the minimal non-zero regression values is fulfilled and $\gamma = 3s/\sqrt{\phi_\infty}$ otherwise. Using $\|\delta\beta_{M^c}\|_1 \leq \|\delta\beta\|_1$ and the bounds in (11) and (12) for the latter quantity, it holds with probability at least $1-p(1-\Phi(C))$ that

$$\|\delta\beta\|_1 \leq \frac{2C\sigma\gamma}{\sqrt{n}}\max\{\frac{s}{\phi_\nu},\frac{4}{\nu}\}.$$

Hence

$$\|\mathbf{X}(\hat{\beta}^{oracle}-\hat{\beta})\|_2^2 = \delta\beta^T\hat{\mathbf{\Sigma}}\delta\beta \;\leq\; 2\frac{(C\sigma\gamma)^2}{n}\max\{\frac{2s}{\phi_\nu},\frac{8}{\nu}\}.$$

Using again $C^2 = K^2 = 2\log(\frac{p}{\sqrt{2\pi\eta}})$, the claim follows.

6.3. Proof of Theorem 3

Proof. The proof follows from equation (15) in the proof of Lemma 1, which is making use of the assumption that $\phi^2_{compatible}(0,S,\hat{\mathbf{\Sigma}}) \geq \phi_\infty$. For any $C>0$, with probability $1-p(1-\Phi(C))$,

$$\delta\beta^T\hat{\mathbf{\Sigma}}\delta\beta \;\leq\; 6\frac{Cs\sigma}{\sqrt{\phi_\infty n}}\|\delta\beta_{M^c}\|_1. \tag{8}$$

□

The vector $\delta\beta$ has by definition positive entries $\delta\beta_k$ for all $k\in N$ as the oracle estimator (5) has identically 0 entries for $k\in N$. Now using the assumption $\phi^2_{pos,S}(\hat{\mathbf{\Sigma}}) \geq \kappa > 0$, the lhs in (8) is greater than or equal to $\kappa\|\delta\beta\|_1^2$. Using $\|\delta\beta_{M^c}\|_1 \leq \|\delta\beta\|_1$ on the rhs, it follows that with probability $1-p(1-\Phi(C))$,

$$\|\delta\beta\|_1 \;\leq\; 6\frac{Cs\sigma}{\kappa\sqrt{\phi_\infty n}}.$$

Hence, as in the proof of Lemma 1, it follows with Lemma 4 that $\|\hat{\beta}-\beta^*\|_1 \leq \|\hat{\beta}^{oracle}-\beta^*\|_1 + \|\delta\beta\|_1$ and hence, with probability $1-p(1-\Phi(C))$, using the same union bound argument as in Lemma 1,

$$\|\hat{\beta}-\beta^*\|_1 \;\leq\; (\frac{6}{\kappa}+2)\frac{Cs\sigma}{\sqrt{\phi_\infty n}} \leq \frac{8Cs\sigma}{\kappa\sqrt{\phi_\infty n}}.$$

Using the same choice of C as in Theorem 1, the proof is complete.

6.4. Lemmata

Lemma 1. *Assume that the* Positive Eigenvalue Condition *holds with $\nu > 0$. Choose any $C > 0$. Assume that the compatibility condition holds with $\phi_\nu > 0$ for $L_\nu = 3/\sqrt{\nu}$ and with $\phi_\infty \geq \phi_\nu$ for $L_\infty = 0$. It then holds with probability at least $1 - p(1 - \Phi(C))$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution,*

$$\|\hat{\beta} - \beta^*\|_1 \ \leq \ \frac{2Cs\sigma}{\sqrt{\phi_\infty n}}\Big(1 + \max\{\frac{3s}{\phi_\nu}, \frac{12}{\nu}\}\Big).$$

If, additionally, the minimal eigenvalue $\phi^2_{min}(\hat{\mathbf{\Sigma}}_{SS})$ is greater than or equal to some $\tau > 0$ and $\min_{k\in S} \beta_k > C\sigma/\sqrt{n\tau}$, then with probability at least $1 - p(1 - \Phi(C))$,

$$\|\hat{\beta} - \beta^*\|_1 \ \leq \ \frac{2Cs\sigma}{\sqrt{\phi_\infty n}}\Big(1 + \max\{\frac{\sqrt{\phi_\infty}}{\phi_\nu}, \frac{4\sqrt{\phi_\infty}}{s\nu}\}\Big).$$

Proof. By definition (1) of $\hat{\beta}$, definition (5) of $\hat{\beta}^{oracle}$ and for $\delta\beta = \hat{\beta} - \hat{\beta}^{oracle}$ we have

$$\begin{aligned} \delta\beta \ &= \ \text{argmin}_\gamma \ \|\mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle} - \mathbf{X}\gamma\|_2^2 \quad \text{such that} \\ \gamma_k &\geq -\hat{\beta}^{oracle}_k \text{ for all } k = 1, \ldots, p. \end{aligned} \tag{9}$$

The bound for $\|\hat{\beta} - \beta^*\|_1$ follows as $\|\hat{\beta} - \beta^*\|_1 = \|\delta\beta + (\hat{\beta}^{oracle} - \beta^*)\|_1 \leq \|\hat{\beta}^{oracle} - \beta^*\|_1 + \|\delta\beta\|_1$. Using Lemma 4, it holds with probability exceeding $1 - 2(1 - \Phi(C))$,

$$\|\hat{\beta}^{oracle} - \beta^*\|_1 \leq \frac{2Cs\sigma}{\sqrt{\phi_\infty n}}, \tag{10}$$

and it thus remains to be shown that, if (10) is fulfilled, it holds with probability at least $1 - p(1 - \Phi(C))$ that

$$\|\delta\beta\|_1 \leq \frac{2Cs\sigma}{\sqrt{\phi_\infty n}} \max\{\frac{3s}{\phi_\nu}, \frac{12}{\nu}\}. \tag{11}$$

Note that the union bound needs a factor p instead of $p + 2$ when combining (10) and (11) as the event (10) was already used in the union bound of (11). If the condition on the minimal non-zero coefficient is fulfilled, the inequality (11) has to hold as

$$\|\delta\beta\|_1 \leq \frac{2C\sigma}{\sqrt{n}} \max\{\frac{s}{\phi_\nu}, \frac{4}{\nu}\}. \tag{12}$$

Let $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle}$. Since $\delta\beta \equiv 0$ is a feasible solution in (9),

$$\delta\beta^T\mathbf{X}^T\mathbf{X}\delta\beta - 2\mathbf{R}^T\mathbf{X}\delta\beta \leq 0.$$

and

$$\delta\beta^T\hat{\mathbf{\Sigma}}\delta\beta \leq \frac{2}{n}\mathbf{R}^T\mathbf{X}\delta\beta. \tag{13}$$

Let

$$M := \{k : \delta\beta_k < 0\} \tag{14}$$

By definition $M \subseteq S$ and $N \subseteq M^c$. The conditions for Lemma 2 are fulfilled, and we thus have with probability at least $1 - p(1 - \Phi(C))$ that

$$\frac{1}{n}\mathbf{R}^T\mathbf{X}\delta\beta \;\le\; \frac{3C\sigma s}{\sqrt{n\phi_\infty}}\|\delta\beta_N\|_1 \;\le\; \frac{3C\sigma s}{\sqrt{n\phi_\infty}}\|\delta\beta_{M^c}\|_1.$$

If additionally the bound on the minimal positive coefficient holds, we get from Lemma 3 that we have instead a bound

$$\frac{1}{n}\mathbf{R}^T\mathbf{X}\delta\beta \;\le\; \frac{C\sigma}{\sqrt{n}}\|\delta\beta_{M^c}\|_1.$$

In either case,

$$\delta\beta^T\hat{\mathbf{\Sigma}}\delta\beta \;\le\; 2\frac{C\sigma\gamma}{\sqrt{n}}\|\delta\beta_{M^c}\|_1, \tag{15}$$

where $\gamma = 1$ if the bound on the minimal non-zero regression values is fulfilled and $\gamma = 3s/\sqrt{\phi_\infty}$ otherwise. Let $a = (\delta\beta_{M^c}\hat{\mathbf{\Sigma}}\delta\beta_{M^c})^{1/2}$ and $b = (\delta\beta_M\hat{\mathbf{\Sigma}}\delta\beta_M)^{1/2}$. Then

$$\begin{aligned}\delta\beta^T\hat{\mathbf{\Sigma}}\delta\beta &\ge a^2 + b^2 - 2ab \\ &\ge a^2 - 2\|\delta\beta_M\|_1 a\end{aligned} \tag{16}$$

having used the normalization of all columns of $\mathbf{X}$ (which bounds the absolute values of all entries in $\hat{\mathbf{\Sigma}}$ by 1) for a bound $b \le \|\delta\beta_M\|_1$. It follows with the *Positive Eigenvalue Condition* for the first term in the last inequality, together with the fact that $\min_{k\in M^c}\delta\beta_k \ge 0$ by definition of M in (14), that $a^2 \ge \nu\|\delta\beta_{M^c}\|_1^2$ for some $0 < \nu \le 1$.

Evidently $\|\delta\beta_{M^c}\|_1 > (3/\sqrt{\nu})\|\delta\beta_M\|_1$ is either true or not. If it is true, the rhs in (16) is then greater than $\nu\|\delta\beta_{M^c}\|_1^2 - 2(\sqrt{\nu}/3)\|\delta\beta_{M^c}\|_1(\sqrt{\nu}\|\delta\beta_{M^c}\|_1)$ and hence greater than $(\nu/3)\|\delta\beta_{M^c}\|_1$. Using this bound on the lhs of (15) and dividing by $\|\delta\beta_{M^c}\|_1$ yields that, with probability at least $1 - p(1 - \Phi(C))$,

$$\|\delta\beta_{M^c}\|_1 \;\le\; \frac{6}{\nu}\frac{C\sigma\gamma}{\sqrt{n}} \tag{17}$$

Together with the assumption for this case that $\|\delta\beta_{M^c}\|_1 > (3/\sqrt{\nu})\|\delta\beta_M\|_1$ and $\nu < 1$, it holds that $\|\delta\beta_M\|_1$ is bounded by $(2/\sqrt{\nu})C\sigma\gamma/\sqrt{n}$. Combining with (17) and $\nu \le 1$,

$$\|\delta\beta\|_1 \;\le\; \frac{8}{\nu}\frac{C\sigma\gamma}{\sqrt{n}}, \tag{18}$$

which satisfies the bounds in (11) and (12).

Assume now the second case, $\|\delta\beta_{M^c}\|_1 \le (3/\sqrt{\nu})\|\delta\beta_M\|_1$. We have then $\|\delta\beta_{M^c}\|_1 \le L_\nu\|\delta\beta_M\|_1$ for $L_\nu = 3/\sqrt{\nu}$ and thus, using $N \subseteq M^c$, also $\|\delta\beta_N\|_1 \le$

$L_\nu \|\delta\beta_S\|_1$. The vector $\delta\beta$ is then in $\mathcal{R}(L_\nu, S)$. Using the compatibility condition, it follows that $\delta\beta^T \hat{\mathbf{\Sigma}} \delta\beta$ is greater than or equal to $(\phi_\nu/s)\|\delta\beta\|_1^2$. Using this on the lhs of (15),

$$\frac{\phi_\nu}{s}\|\delta\beta\|_1^2 \;\le\; 2\frac{C\sigma\gamma}{\sqrt{n}}\|\delta\beta_{M^c}\|_1. \tag{19}$$

Since $\|\delta\beta_{M^c}\|_1 = \|\delta\beta\|_1 - \|\delta\beta_M\|_1 \le \|\delta\beta\|_1$, it follows that

$$\|\delta\beta\|_1 \le \frac{2s}{\phi_\nu}\frac{C\sigma\gamma}{\sqrt{n}}, \tag{20}$$

which also satisfies the bounds in (11) and (12) . Hence, the bounds (11) and (12) hold under both possible scenarios ($\|\delta\beta_{M^c}\|_1 > (3/\sqrt{\nu})\|\delta\beta_M\|_1$ true or false) and the proof is complete. □

Lemma 2. *Let $\delta\beta = \hat{\beta} - \hat{\beta}^{oracle}$ and $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle}$ with $\hat{\beta}^{oracle}$ defined as in (5). Assume that $\phi^2_{compatible}(0, S, \hat{\mathbf{\Sigma}}) \ge \phi_\infty > 0$, With probability at least $1 - p(1 - \Phi(C))$,*

$$\mathbf{R}^T\mathbf{X}\delta\beta \;\le\; 3C\sigma s\sqrt{\frac{n}{\phi_\infty}}\|\delta\beta_N\|_1.$$

Proof. First, we can write

$$\mathbf{R}^T\mathbf{X}\delta\beta \;=\; \sum_{k\in S}(\mathbf{R}^T\mathbf{X}_k)\delta\beta_k + \sum_{k\in N}(\mathbf{R}^T\mathbf{X}_k)\delta\beta_k \tag{21}$$

By definition of $\hat{\beta}^{oracle}$ and using the KKT conditions, we have for all $k \in S$ that either (a) $\hat{\beta}_k^{oracle} > 0$ and $\mathbf{R}^T\mathbf{X}_k = 0$ or (b) $\hat{\beta}_k^{oracle} = 0$ and $\mathbf{R}^T\mathbf{X}_k \le 0$. The contribution from all (a) cases vanishes in (21). For $k \in S$ that fall into category (b), it follows by $\hat{\beta}_k \ge 0$ and $\hat{\beta}_k^{oracle} = 0$ that $\delta\beta_k \ge 0$ and hence $(\mathbf{R}^T\mathbf{X}_k)\delta\beta_k \le 0$. We are left with contributions from $k \in N$ in (21),

$$\mathbf{R}^T\mathbf{X}\delta\beta \;\le\; \sum_{k\in N}(\mathbf{R}^T\mathbf{X}_k)\delta\beta_k \;\le\; \max_{k\in N}(\mathbf{R}^T\mathbf{X}_k)\|\delta\beta_N\|_1 \tag{22}$$

It thus remains to be shown that, with probability at least $1 - p(1 - \Phi(C))$,

$$\max_{k\in N}(\mathbf{R}^T\mathbf{X}_k) \le 3C\sigma s\sqrt{n/\phi_\infty} \tag{23}$$

To show this, write $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle} = \varepsilon + \mathbf{X}(\beta^* - \hat{\beta}^{oracle})$. Then

$$\max_{k\in N}\mathbf{R}^T\mathbf{X}_k \;\le\; \max_{k\in N}\varepsilon^T\mathbf{X}_k + \|\beta^* - \hat{\beta}^{oracle}\|_1 \max_{k'\in S, k\in N}\mathbf{X}_{k'}^T\mathbf{X}_k. \tag{24}$$

Taking a union bound yields that, with probability at least $1-(p-2)(1-\Phi(C))$, $\varepsilon^T\mathbf{X}_k \le \sigma\sqrt{n}$ for all $k \in N$ (as $|N| \le p-2$), having used the normalisations of the columns in $\mathbf{X}$. Using the same normalisation, the second term on the rhs in

(24) is bounded by $n\|\beta^* - \hat{\beta}^{oracle}\|_1$. This contribution is bounded with probability $1 - 2(1 - \Phi(C))$ by $2Cs\sigma/\sqrt{n\phi_\infty}$ with Lemma 4. Hence, with probability at least $1 - p(1 - \Phi(C))$,

$$\max_{k\in N} \mathbf{R}^T\mathbf{X}_k \;\le\; \sigma\sqrt{n} + 2\sigma sC\sqrt{n/\phi_\infty} = \sigma\sqrt{n}(1 + 2sC/\sqrt{\phi_\infty}) \le 3C\sigma s\sqrt{n/\phi_\infty}, \tag{25}$$

which shows (23) and hence completes the proof. □

Lemma 3. *Let $\delta\beta = \hat{\beta} - \hat{\beta}^{oracle}$ and $\mathbf{R} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle}$ with $\hat{\beta}^{oracle}$ defined as in (5). If the minimal eigenvalue $\phi^2_{min}(\hat{\mathbf{\Sigma}}_{SS})$ is greater than or equal to some $\tau > 0$ and $\min_{k\in S}\beta_k > C\sigma/\sqrt{n\tau}$, with probability at least $1 - p(1 - \Phi(C))$,*

$$\mathbf{R}^T\mathbf{X}\delta\beta \;\le\; C\sigma\sqrt{n}\|\delta\beta_N\|_1.$$

Proof. The proof is analogous to the proof of Lemma 2 in the beginning and we are again left with contributions from $k \in N$ in (21),

$$\mathbf{R}^T\mathbf{X}\delta\beta \;\le\; \sum_{k\in N}(\mathbf{R}^T\mathbf{X}_k)\delta\beta_k \;\le\; \max_{k\in N}(\mathbf{R}^T\mathbf{X}_k)\|\delta\beta_N\|_1 \tag{26}$$

It thus remains to be shown that, with probability at least $1 - p(1 - \Phi(C))$,

$$\max_{k\in N}(\mathbf{R}^T\mathbf{X}_k) \le C\sigma\sqrt{n} \tag{27}$$

and this follows from Lemma 6, which completes the proof. □

Lemma 4. *If $\phi^2_{compatible}(0, S, \hat{\mathbf{\Sigma}}) \ge \phi_\infty$, with probability at least $1-2(1-\Phi(C))$,*

$$\|\beta^* - \hat{\beta}^{oracle}\|_1 \;\le\; 2C\sigma\frac{s}{\sqrt{n\phi_\infty}}.$$

Proof. Let P_S be the projection into the space spannend by the columns in $\mathbf{X}_S$. By definition of $\hat{\beta}^{oracle}$, the vector β^* is a feasible solution in the optimization problem of $\hat{\beta}^{oracle}$. As $\beta^*_k = \hat{\beta}^{oracle}_k = 0$ for all $k \in N = S^c$,

$$\|P_S\mathbf{Y} - \mathbf{X}\beta^*\|_2^2 \ge \|P_S\mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle}\|_2^2.$$

Furthermore,

$$\|P_S\mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle}\|_2^2 \ge \left(\|\mathbf{X}\beta^* - \mathbf{X}\hat{\beta}^{oracle}\|_2 - \|P_S\mathbf{Y} - \mathbf{X}\beta^*\|_2\right)^2.$$

Putting the last two equations together,

$$\|\mathbf{X}(\beta^* - \hat{\beta}^{oracle})\|_2 \le 2\|P_S\mathbf{Y} - \mathbf{X}\beta^*\|_2 \tag{28}$$

Since $\beta^* - \hat{\beta}^{oracle}$ vanishes identically in $N = S^c$, the vector is in $\mathcal{R}(0, S)$ and the lhs in (28) is larger than $\sqrt{\phi_\infty n/s}\|\beta^* - \hat{\beta}^{oracle}\|_1$. On the rhs, $\sigma^{-2}\|P_S\mathbf{Y} - \mathbf{X}\beta^*\|_2^2 = \sigma^{-2}\|P_S\varepsilon\|_2^2$ has a χ^2_s-distribution. Then $\sigma^{-2}\|P_S\mathbf{Y} - \mathbf{X}\beta^*\|_2^2 > sC$ with

probability at most $2\tilde{\Phi}(C)$ for all s, where $\tilde{\Phi}(\cdot) = 1 - \Phi(\cdot)$. Together, it follows that with probability at least $1 - 2(1 - \Phi(C))$,

$$\|\beta^* - \hat{\beta}^{oracle}\|_1 \leq 2C\sigma \frac{s}{\sqrt{n\phi_\infty}},$$

which completes the proof. □

Lemma 5. *Let $\hat{\beta}^{ols}$ be the least squares estimator restricted to S:*

$$\hat{\beta}^{ols} \;=\; argmin_\beta \, \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \;\; such \; that \;\; \beta_N \equiv 0.$$

If the minimal eigenvalue $\phi^2_{min}(\hat{\mathbf{\Sigma}}_{SS})$ is greater than or equal to some $\tau > 0$ and $\min_{k\in S} \beta_k > C\sigma/\sqrt{n\tau}$, it holds that $P(\hat{\beta}^{ols} \equiv \hat{\beta}^{oracle}) \geq 1 - s(1 - \Phi(C))$. With at least the same probability $1 - s(1 - \Phi(C))$,

$$\|\beta^* - \hat{\beta}^{oracle}\|_\infty \;\leq\; C\sigma/\sqrt{n\tau}$$

Proof. It is only necessary to show that $\min_{k\in S} \hat{\beta}^{ols}_k \geq 0$ with probability at least $1 - s(1 - \Phi(C))$.

The error term has, under the made assumptions, a normal distribution, $\hat{\beta}^{ols}_k - \beta^* \sim \mathcal{N}(0, \sigma^2(n\hat{\mathbf{\Sigma}}_{SS})^{-1}_k)$ for all $k \in S$. The minimal eigenvalue of $\hat{\mathbf{\Sigma}}_{SS}$ is bounded from below by τ by the made assumption and the variance of $\hat{\beta}^{ols}_k$ is thus bounded from above by $\sigma^2/(n\tau)$ for all $k \in S$. It follows with Bonferroni's inequality that, with probability at least $1 - s(1 - \Phi(C))$,

$$\|\beta^* - \hat{\beta}^{ols}\|_\infty \leq C\sigma/\sqrt{n\tau}. \tag{29}$$

If $\min_{k\in S} \beta^*_k \geq C\sigma/\sqrt{n\tau}$, then (29) implies that $\min_{k\in S} \hat{\beta}^{ols}_k \geq 0$ and thus $\hat{\beta}^{oracle} \equiv \hat{\beta}^{ols}$ and thus also

$$\|\beta^* - \hat{\beta}^{oracle}\|_\infty \leq C\sigma/\sqrt{n\tau},$$

which completes the proof. □

Lemma 6. *If the minimal eigenvalue $\phi^2_{min}(\hat{\mathbf{\Sigma}}_{SS})$ is greater than or equal to some $\tau > 0$ and $\min_{k\in S} \beta_k > C\sigma/\sqrt{n\tau}$, with probability at least $1-p(1-\Phi(C))$,*

$$\max_{k\in N} (\mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle})^T \mathbf{X}_k \leq C\sigma\sqrt{n}.$$

Proof. We condition on the event $\hat{\beta}^{oracle} \equiv \hat{\beta}^{ols}$, which happens according to Lemma 5 with probability at least $1 - s(1 - \Phi(C))$. Then $\mathbf{Y} - \mathbf{X}\hat{\beta}^{oracle} = \mathbf{Y} - \mathbf{X}\hat{\beta}^{ols} = P_{S\perp}\mathbf{Y}$, where $P_{S\perp}\mathbf{Z}$ is the projection of a vector $\mathbf{Z} \in \mathbb{R}^n$ into the space orthogonal to $\mathbf{X}_S$. Now, $P_{S\perp}\mathbf{Y} = P_{S\perp}(\mathbf{X}\beta^* + \varepsilon) = P_{S\perp}\varepsilon$. The distribution of $(P_{S\perp}\varepsilon)^T\mathbf{X}_k$ is, for every $k \in N$, normal with mean 0 and variance at most $\sigma^2 n$, and thus $P((P_{S\perp}\varepsilon)^T\mathbf{X}_k \geq C\sigma\sqrt{n}) \leq 1 - \Phi(C)$ for all $k \in N$ and, using a Bonferroni bound, $P(\max_{k\in N}(P_{S\perp}\varepsilon)^T\mathbf{X}_k \geq C\sigma\sqrt{n}) \leq |N|(1 - \Phi(C))$. The unconditional probability of $\max_{k\in N}(P_{S\perp}\varepsilon)^T\mathbf{X}_k \geq C\sigma\sqrt{n}$ is thus at least $1 - s(1-\Phi(C)) - |N|(1-\Phi(C)) = 1-(s+|N|)(1-\Phi(C)) = 1-p(1-\Phi(C))$, which completes the proof. □

References

Bellavia, S., Macconi, M., and Morini, B., An interior point newton-like method for non-negative least-squares problems with degenerate solution. *Numerical Linear Algebra with Applications*, 13:825–846, 2006. MR2278195

Belloni, A., Chernozhukov, V., and Wang, L., Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98:791–806, 2011. MR2860324

Bickel, P., Ritov, Y., and Tsybakov, A., Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009. MR2533469

Boutsidis, C. and Drineas, P., Random projections for the nonnegative least-squares problem. *Linear Algebra and its Applications*, 431:760–771, 2009. MR2535548

Breiman, L., Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995. MR1365720

Bruckstein, A.M., Elad, M., and Zibulevsky, M., On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54:4813–4820, 2008. MR2589866

Bunea, B.F., Tsybakov, A.B., and Wegkamp, M.H., Aggregation for gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007a. MR2351101

Bunea, F., Tsybakov, A., and Wegkamp, M., Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007b. MR2312149

Bunea, F., Lederer, J., and She, Y., The group square-root lasso: Theoretical properties and fast algorithms. *arXiv preprint arXiv:1302.0261*, 2013.

Candes, E. and Tao, T., The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2312–2351, 2007. MR2382644

Castro, R., Coates, M., Liang, G., Nowak, R., and Yu, B., Network tomography: Recent developments. *Statistical Science*, 3:499–517, 2004. MR2185628

Chen, D. and Plemmons, R.J., *Nonnegativity constraints in numerical analysis*. World Scientific Press, River Edge, NJ, USA, 2009. MR2604144

Ding, C.H.Q., Li, T., and Jordan, M.I., Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:45–55, 2010.

Donoho, D.L., Johnstone, I.M., Hoch, J.C., and Stern, A.S., Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B*, 54:41–81, 1992. MR1157714

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., Least angle regression. *Annals of Statistics*, 32:407–451, 2004. MR2060166

Guo, Yi and Berman, Mark, A comparison between subset selection and l1 regularisation with an application in spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 118:127–138, 2012.

Hoerl, Arthur E. and Kennard, Robert W., Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Kim, D., Sra, S., and Dhillon, I.S., A new projected quasi-newton approach for the nonnegative least squares problem. Technical report, Department of Computer Science, University of Texas, TR-06-54, 2006.

Kim, H. and Park, H., Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23:1495, 2007.

Lawrence, E., Michailidis, G., and Nair, V.N., Network delay tomography using flexicast experiments. *Journal of the Royal Statistical Society: Series B*, 68:785–813, 2006. MR2301295

Lawson, C.L. and Hanson, R.J., *Solving least squares problems*, volume 15. Society for Industrial Mathematics, 1995. MR1349828

Lee, D.D. and Seung, H.S., Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.

Lee, D.D., Seung, H.S., et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

Meinshausen, N., Relaxed lasso. *Computational Statistics and Data Analysis*, 52:374–393, 2007. MR2409990

Meinshausen, N. and Yu, B., Lasso-type recovery of sparse representations from high-dimensional data. *Annals of Statistics*, 7:246–270, 2009. MR2488351

Slawski, M. and Hein, M., Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization. *arXiv preprint arXiv:1205.0953*, 2012.

Slawski, M., Hein, M., and Campus, E., Sparse recovery by thresholded non-negative least squares. Technical report, Department of Computer Science, University of Saarbruecken, 2011.

Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. MR1379242

Van De Geer, S.A., High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008. MR2396809

Van De Geer, S.A. and Bühlmann, P., On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. MR2576316

Wainwright, M.J., Sharp thresholds for high-dimensional and noisy recovery of sparsity. *IEEE Transactions on Information Theory*, 55:2183–2202, 2009. MR2729873

Waterman, M.S., Least squares with nonnegative regression coefficients. *Journal of Statistical Computation and Simulation*, 6:67–70, 1977.

Xi, B., Michailidis, G., and Nair, V.N., Estimating network loss rates using active tomography. *Journal of the American Statistical Association*, 101:1430–1448, 2006. MR2279470

Yuan, M. and Lin, Y., Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68:49–67, 2006. MR2212574

Zhang, C.H. and Huang, J., The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008. MR2435448

Zou, H., The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. MR2279469