Electronic Journal of Statistics
Vol. 7 (2013) 3004–3056
ISSN: 1935-7524
DOI: 10.1214/13-EJS868

Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization

Martin Slawski* **and Matthias Hein**

Saarland University
Campus E 1.1
Saarbrücken, Germany
e-mail: ms@cs.uni-saarland.de; hein@cs.uni-saarland.de

Abstract: Least squares fitting is in general not useful for high-dimensional linear models, in which the number of predictors is of the same or even larger order of magnitude than the number of samples. Theory developed in recent years has coined a paradigm according to which sparsity-promoting regularization is regarded as a necessity in such setting. Deviating from this paradigm, we show that non-negativity constraints on the regression coefficients may be similarly effective as explicit regularization if the design matrix has additional properties, which are met in several applications of non-negative least squares (NNLS). We show that for these designs, the performance of NNLS with regard to prediction and estimation is comparable to that of the lasso. We argue further that in specific cases, NNLS may have a better ℓ_∞-rate in estimation and hence also advantages with respect to support recovery when combined with thresholding. From a practical point of view, NNLS does not depend on a regularization parameter and is hence easier to use.

AMS 2000 subject classifications: Primary 62J05; secondary 52B99.
Keywords and phrases: Convex geometry, deconvolution, high dimensions, non-negativity constraints, persistence, random matrices, separating hyperplane, sparse recovery.

Received March 2013.

1. Introduction

Consider the linear regression model

$$y = X\beta^* + \varepsilon, \tag{1.1}$$

where y is a vector of observations, $X \in \mathbb{R}^{n\times p}$ a design matrix, ε a vector of noise and β^* a vector of coefficients to be estimated. Throughout this paper, we are concerned with a high-dimensional setting in which the number of unknowns p is at least of the same order of magnitude as the number of observations n,

*Supported by the Cluster of Excellence 'Multimodal Computing and Interaction' (MMCI) of Deutsche Forschungsgemeinschaft (DFG).

i.e. $p = \Theta(n)$ or even $p \gg n$, in which case one cannot hope to recover the target β^* if it does not satisfy one of various kinds of sparsity constraints, the simplest being that β^* is supported on $S = \{j : \beta_j^* \neq 0\}$, $|S| = s < n$. In this paper, we additionally assume that β^* is non-negative, i.e. $\beta^* \in \mathbb{R}_+^p$. This constraint is particularly relevant when modelling non-negative data, which emerge e.g. in the form of pixel intensity values of an image, time measurements, histograms or count data, economical quantities such as prices, incomes and growth rates. Non-negativity constraints occur naturally in numerous deconvolution and unmixing problems in diverse fields such as acoustics [31], astronomical imaging [2], hyperspectral imaging [1], genomics [30], proteomics [42], spectroscopy [16] and network tomography [34]; see [14] for a survey. As reported in these references, non-negative least squares (NNLS) yields at least reasonably good, sometimes even excellent results in practice, which may seem surprising in view of the simplicity of that approach. The NNLS problem is given by the quadratic program

$$\min_{\beta \succeq 0} \frac{1}{n} \|y - X\beta\|_2^2, \tag{1.2}$$

which is a convex optimization problem that can be solved efficiently [27]. A minimizer $\widehat{\beta}$ of (1.2) will be referred to as an NNLS estimator. Solid theoretical support for the empirical success of NNLS from a statistical perspective scarcely appears in the literature. An early reference is [16] dating back already two decades. The authors show that, depending on X and the sparsity of β^*, NNLS may have a 'super-resolution'-property that permits reliable estimation of β^*. Rather recently, sparse recovery of non-negative signals in a noiseless setting ($\varepsilon = 0$) has been studied in [7, 19, 55, 56]. One important finding of this body of work is that non-negativity constraints alone may suffice for sparse recovery, without the need to employ sparsity-promoting ℓ_1-regularization as usually. On the other hand, it remains unclear whether similar results continue to hold in a more realistic noisy setup. At first glance, the following considerations suggest a negative answer. The fact that NNLS, apart from sign constraints, only consists of a fitting term, fosters the intuition that NNLS is prone to overfitting, specifically in a high-dimensional setting. Usefulness of NNLS in such setting appears to be in conflict with the well-established paradigm that a regularizer is necessary to prevent over-adaptation to noise and to enforce desired structural properties of the solution, like sparsity of the vector of coefficients. As one of the main contributions of the present paper, we characterize the *self-regularizing property* which NNLS exhibits for a certain class of designs that turn out to be tailored to the non-negativity constraint, thereby disentangling the apparent conflict above and improving the understanding of the empirical success of NNLS. More precisely, we show that for these designs, NNLS is rather closely related to the non-negative lasso problem

$$\min_{\beta \succeq 0} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \mathbf{1}^\top \beta, \quad \lambda > 0, \tag{1.3}$$

the sign-constrained counterpart to the popular lasso problem [47], which is given by

$$\min_{\beta \geq 0} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1, \quad \lambda > 0, \tag{1.4}$$

where $\lambda > 0$ is a tuning parameter. Elaborating on the relation between NNLS and the non-negative lasso, we establish that for the aforementioned class of designs, NNLS achieves comparable performance with regard to prediction, estimation of β^* and support recovery when combined with subsequent thresholding. We here refer to the monograph [8], the survey [52] and the retrospective [48] for an overview on the appealing performance guarantees established for the lasso in the last decade, which have, along with favourable computational properties, contributed to its enormous popularity in data analysis. In this paper, we argue that, in view of both theoretical considerations and empirical evidence, improvements of NNLS over the (non-negative) lasso are possible, even though they are limited to a comparatively small set of designs. Differences in performance arise from the bias of the ℓ_1-regularizer in (1.3) and (1.4) that is responsible for a in general sub-optimal rate for estimation of β^* in the ℓ_∞-norm [60]. Unlike for NNLS, a tuning parameter needs to be specified for the (non-negative) lasso, as it is necessary for all regularization based-methods. Selection of the tuning parameter by means of cross-validation increases the computational burden and may be error-prone if done by a grid search, since the grid could have an unfavourable range or a too small number of grid points. Theoretical results on how to set the regularization parameter are often available, but require a sufficient degree of acquaintance with existing literature and possibly also knowledge of quantities such as the noise level. The last issue has been withstanding problem of the lasso until recently. In [4] and [46] two related variants of the lasso are proposed that have similar theoretical guarantees, while the tuning parameter can be set without knowledge of the noise level. On the other hand, NNLS is directly applicable, since it is free of tuning parameters.

Outline and contributions of the paper. The paper significantly extends a previous conference publication [43], which contains the first systematic analysis of NNLS in a high-dimensional setting. Recently, Meinshausen [34] has independently studied the performance of NNLS in such a setting. His work is related to ours in Section 3. The paper is organized as follows. In Section 2 we work out the self-regularizing property NNLS may have in conjunction with certain design matrices. Equipped with that property, a bound on the prediction error is stated that resembles a corresponding 'slow rate' bound available for the lasso. Developing further the connection to the lasso, we use techniques pioneered in Bickel *et al.* [5] to prove bounds on the estimation error $\|\widehat{\beta} - \beta^*\|_q$ in ℓ_q-norm, $q \in [1, 2]$, and an improved bound on the prediction error of NNLS in Section 3. In Section 4, we finally provide bounds on the sup-norm error $\|\widehat{\beta} - \beta^*\|_\infty$ of NNLS. Hard thresholding of $\widehat{\beta}$ is proposed for sparse recovery, and a data-driven procedure for selecting the threshold is devised. Section 4 also contains a discussion of advantages and disadvantages of NNLS relative to the non-negative lasso. In Section 5, we have a closer look at several designs, which

satisfy the conditions required throughout the paper. In Section 6, we discuss the empirical performance of NNLS in deconvolution and sparse recovery in comparison to standard methods, in particular to the non-negative lasso. The appendix contains most of the proofs, apart from those that have been placed into a supplementary file [45] for space reasons.

Notation. We denote the usual ℓ_q-norm by $\|\cdot\|_q$. The cardinality of a set is denoted by $|\cdot|$. Let $J, K \subseteq \{1, \dots, m\}$ be index sets. For a matrix $A \in \mathbb{R}^{n \times m}$, A_J denotes the matrix one obtains by extracting the columns corresponding to J. For $j = 1, \dots, m$, A_j denotes the j-th column of A. The matrix A_{JK} is the sub-matrix of A by extracting rows in J and columns in K. Likewise, for $v \in \mathbb{R}^m$, v_J is the sub-vector corresponding to J. Identity matrices and vectors of ones of varying dimensions are denoted by I respectively $\mathbf{1}$. The symbols $\preceq, \succeq$ and $\prec, \succ$ denote componentwise inequalities and componentwise strict inequalities, respectively. In addition, for some matrix A, $A \succeq a$ means that all entries of A are at least equal to a scalar a. The non-negative orthant $\{x \in \mathbb{R}^m : x \succeq 0\}$ is denoted by $\mathbb{R}^m_+$. The standard simplex in $\mathbb{R}^m$, that is the set $\{x \in \mathbb{R}^m_+ : \sum_{j=1}^m x_j = 1\}$ is denoted by T^{m-1}. Lower and uppercase c's like c, c', c_1 and C, C', C_1 etc. denote positive constants not depending on the sample size n whose values may differ from line to line. In general, the positive integers $p = p_n$ and $s = s_n$ depend on n. Landau's symbols are denoted by $o(\cdot), O(\cdot), \Theta(\cdot), \Omega(\cdot)$. Asymptotics is to be understood w.r.t. a triangular array of observations $\{(X_n, y_n), X_n \in \mathbb{R}^{n \times p_n}\}$, $n = 1, 2, \dots$.

Normalization. If not stated otherwise, the design matrix X is considered as deterministic, having its columns normalized such that $\|X_j\|_2^2 = n$, $j = 1, \dots, p$.

General linear position. We say that the columns of X are in general linear position in $\mathbb{R}^n$ if the following condition (GLP) holds

$$\text{(GLP)}: \quad \forall J \subset \{1, \dots, p\},\ |J| = \min\{n, p\}\ \ \forall \lambda \in \mathbb{R}^{|J|}\ \ X_J \lambda = 0 \implies \lambda = 0, \tag{1.5}$$

where for $J \subseteq \{1, \dots, p\}$, X_J denotes the submatrix of X consisting of the columns corresponding to J.

2. Prediction error: A bound for 'self-regularizing' designs

The main result of the following section is a bound on the excess error of NNLS that resembles the so-called slow rate bound of the lasso [3, 8, 24]. In contrast to the latter, the corresponding bound for NNLS only holds for a certain class of designs. We first show that extra conditions on the design are in fact necessary to obtain such bound. We then introduce a condition on X that we refer to as 'self-regularizing property' which is sufficient to establish a slow rate bound for NNLS. The term 'self-regularization' is motivated from a resulting decomposition of the least squares objective into a modified fitting term and a quadratic term that plays a role similar to an ℓ_1-penalty on the coefficients. This finding provides an intuitive explanation for the fact that NNLS may achieve similar performance than the lasso, albeit no explicit regularization is employed.

2.1. A minimum requirement on the design for non-negativity being an actual constraint

In general, the non-negativity constraints in (1.2) may not be meaningful at all, given the fact that any least squares problem can be reformulated as a non-negative least squares problem with an augmented design matrix $[X\ -X]$. More generally, NNLS can be as ill-posed as least squares if the following condition $(\mathcal{H})$ does not hold.

$$(\mathcal{H}):\quad \exists w \in \mathbb{R}^n \text{ such that } X^\top w \succ 0. \tag{2.1}$$

Condition $(\mathcal{H})$ requires the columns of X be contained in the interior of a half-space containing the origin. If $(\mathcal{H})$ fails to hold, $0 \in \mathrm{conv}\{X_j\}_{j=1}^p$, so that there are infinitely many minimizers of the NNLS problem (1.2). If additionally $p > n$ and the columns of X are in general linear position (condition (GLP) in (1.5)), 0 must be in the interior of $\mathrm{conv}\{X_j\}_{j=1}^p$. It then follows that $\mathcal{C} = \mathbb{R}^n$, where $\mathcal{C} = \mathrm{cone}\{X_j\}_{j=1}^p$ denotes the polyhedral cone generated by the columns of X. Consequently, the non-negativity constraints become vacuous and NNLS yields perfect fit for any observation vector y. In light of this, $(\mathcal{H})$ constitutes a necessary condition for a possible improvement of NNLS over ordinary least squares.

2.2. Overfitting of NNLS for orthonormal design

Since NNLS is a pure fitting approach, over-adaptation to noise is a natural concern. Resistance to overfitting can be quantified in terms of $\frac{1}{n}\|X\widehat{\beta}\|_2^2$ when $\beta^* = 0$ in (1.1). It turns out that condition $(\mathcal{H})$ is not sufficient to ensure that $\frac{1}{n}\|X\widehat{\beta}\|_2^2 = o(1)$ with high probability, as can be seen from studying orthonormal design, i.e. $X^\top X = nI$. Let $y = \varepsilon$ be a standard Gaussian random vector. The NNLS estimator has the closed form expression

$$\widehat{\beta}_j = \max\{X_j^\top \varepsilon, 0\}/n,\quad j = 1, \ldots, p,$$

so that the distribution of each component of $\widehat{\beta}$ is given by a mixture of a point mass 0.5 at zero and a half-normal distribution. We conclude that $\frac{1}{n}\|X\widehat{\beta}\|_2^2 = \frac{1}{n}\|\widehat{\beta}\|_2^2$ is of the order $\Omega(p/n)$ with high probability. The fact that X is orthonormal is much stronger than the obviously necessary half-space constraint $(\mathcal{H})$. In fact, as rendered more precisely in Section 5, orthonormal design turns out to be at the edge of the set of designs still leading to overfitting.

2.3. A sufficient condition on the design preventing NNLS from overfitting

We now present a sufficient condition X has to satisfy so that overfitting is prevented. That condition arises as a direct strengthening of $(\mathcal{H})$. In order to

quantify the separation required in $(\mathcal{H})$, we define

$$\tau_0 = \left\{ \max \tau : \ \exists w \in \mathbb{R}^n, \ \|w\|_2 \leq 1 \quad \text{s.t.} \quad \frac{X^\top w}{\sqrt{n}} \succeq \tau \mathbf{1} \right\}. \tag{2.2}$$

Note that $\tau_0 > 0$ if and only if $(\mathcal{H})$ is fulfilled. Also note that with $\|X_j\|_2^2 = n \ \forall j$, it holds that $\tau_0 \leq 1$. Introducing the Gram matrix $\Sigma = \frac{1}{n} X^\top X$, we have by convex duality that

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \frac{1}{n} \|X\lambda\|_2^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda, \quad \text{where } T^{p-1} = \{\lambda \in \mathbb{R}_+^p : \ \mathbf{1}^\top \lambda = 1\}, \tag{2.3}$$

i.e. in geometrical terms, τ_0 equals the distance of the convex hull of the columns of X (scaled by $1/\sqrt{n}$) to the origin. Using terminology from support vector machine classification (e.g. [41], Sec. 7.2), τ_0 can be interpreted as margin of a maximum margin hyperplane with normal vector w separating the columns of X from the origin. As argued below, in case that τ_0 scales as a constant, overfitting is curbed. This is e.g. *not* fulfilled for orthonormal design, where $\tau_0 = 1/\sqrt{p}$ (cf. Section 5).

Condition 1. A design X is said to have a **self-regularizing property** if there exists a constant $\tau > 0$ so that with τ_0 as defined in (2.2), it holds that $\tau_0 \geq \tau > 0$.

The term 'self-regularization' expresses the fact that the design itself automatically generates a regularizing term, as emphasized in the next proposition and the comments that follow. We point out that Proposition 1 is a qualitative statement preliminary to the main result of the section (Theorem 1) and mainly serves an illustrative purpose.

Proposition 1. *Consider the linear model (1.1) with $\beta^* = 0$ and $y = \varepsilon$ having entries that are independent random variables with zero mean and finite variance. Suppose that X satisfies Condition 1. We then have*

$$\min_{\beta \succeq 0} \frac{1}{n} \|\varepsilon - X\beta\|_2^2 = \min_{\beta \succeq 0} \frac{1}{n} \|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2 (\mathbf{1}^\top \beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{2.4}$$

with $\widetilde{X} = (\Pi X)D$, where Π is a projection onto an $(n-1)$-dimensional subspace of $\mathbb{R}^n$ and D is a diagonal matrix, the diagonal entries being contained in $[\tau, 1]$. Moreover, if $\frac{1}{n}\|X^\top \varepsilon\|_\infty = o_{\mathbf{P}}(1)$, then any minimizer $\widehat{\beta}$ of (2.4) obeys $\frac{1}{n}\|X\widehat{\beta}\|_2^2 = o_{\mathbf{P}}(1)$.

In Proposition 1, the pure noise fitting problem is decomposed into a fitting term with modified design matrix $\widetilde{X}$, a second term that can be interpreted as *squared* non-negative lasso penalty $\tau^2(\mathbf{1}^\top \beta)^2$ (cf. (1.3)) and an additional stochastic term of lower order. As made precise in the proof, the lower bound on τ implies that the ℓ_1-norm of any minimizer is upper bounded by a constant. Prevention of overfitting is then an immediate consequence under the further assumption that the term $\frac{1}{n}\|X^\top \varepsilon\|_\infty = o_{\mathbf{P}}(1)$ tends to zero. This holds under

rather mild additional conditions on X [33] or more stringent conditions on the tails of the noise distribution. As a last comment, let us make the connection of the r.h.s. of (2.4) to a non-negative lasso problem more explicit. Due to the correspondence of the level sets of the mappings $\beta \mapsto \mathbf{1}^\top \beta$ and $\beta \mapsto (\mathbf{1}^\top \beta)^2$ on $\mathbb{R}^p_+$, we have

$$\min_{\beta \succeq 0} \frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2(\mathbf{1}^\top \beta)^2 = \min_{\beta \succeq 0} \frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \gamma(\tau)\mathbf{1}^\top \beta, \tag{2.5}$$

where γ is a non-negative, monotonically increasing function of τ. Proposition 1 in conjunction with (2.5) provides a high-level understanding of what will be shown in the sequel, namely that NNLS may inherit desirable properties of the (non-negative) lasso with regard to prediction, estimation and sparsity of the solution.

2.4. Slow rate bound

Condition 1 gives rise to the following general bound on the ℓ_2-prediction error of NNLS. Note that in Theorem 1 below, it is not assumed that the linear model is specified correctly. Instead, we only assume that there is a fixed target $f = (f_1, \ldots, f_n)^\top$ to be approximated by a non-negative combination of the columns of X.

Theorem 1. *Let $y = f + \varepsilon$, where $f \in \mathbb{R}^n$ is fixed and ε has i.i.d. zero-mean sub-Gaussian entries with parameter σ* [1]*. Define*

$$\mathcal{E}^* = \min_{\beta \succeq 0} \frac{1}{n}\|X\beta - f\|_2^2, \qquad \widehat{\mathcal{E}} = \frac{1}{n}\|X\widehat{\beta} - f\|_2^2.$$

Suppose that X satisfies Condition 1. Then, for any minimizer $\widehat{\beta}$ of the NNLS problem (1.2) and any $M \geq 0$, it holds with probability no less than $1 - 2p^{-M^2}$ that

$$\widehat{\mathcal{E}} \leq \mathcal{E}^* + \left(\frac{6\|\beta^*\|_1 + 8\sqrt{\mathcal{E}^*}}{\tau^2}\right)(1+M)\sigma\sqrt{\frac{2\log p}{n}} + \frac{16(1+M)^2\sigma^2}{\tau^2}\frac{\log p}{n}, \tag{2.6}$$

for all $\beta^* \in \operatorname{argmin}_{\beta \succeq 0} \frac{1}{n}\|X\beta - f\|_2^2$.

Comparison with the slow rate bound of the lasso. Theorem 1 bounds the excess error by a term of order $O(\|\beta^*\|_1 \sqrt{\log(p)/n})$, which implies that NNLS can be consistent in a regime in which the number of predictors p is nearly exponential in the number of observations n. That is, NNLS constitutes a 'persistent procedure' in the spirit of Greenshtein and Ritov [24] who coined the notion of 'persistence' as distinguished from classical consistency with a fixed number of predictors. The excess error bound of Theorem 1 is of the same

[1] See Appendix A for background on sub-Gaussian random variables

order of magnitude as the corresponding bound of the lasso [3, 8, 24] that is typically referred to as slow rate bound. Since the bound (2.6) depends on τ, it is recommended to solve the quadratic program in (2.3) before applying NNLS, which is roughly of the same computational cost. Unlike Theorem 1, the slow rate bound of the lasso does not require any conditions on the design and is more favourable than (2.6) regarding the constants. In [25, 53], improvements of the slow rate bound are derived. On the other hand, the results for the lasso require the regularization parameter to be chosen appropriately.

Remark 1. NNLS has been introduced as a tool for 'non-negative data'. In this context, the assumption of zero-mean noise in Theorem 1 is questionable. In case that the entries of ε have a positive mean, one can decompose ε into a constant term, which can be absorbed into the linear model, and a second term which has mean zero, so that Theorem 1 continues to be applicable.

3. Fast rate bound for prediction and bounds on the ℓ_q-error for estimation, $1 \leq q \leq 2$

Within this section, we further elaborate on the similarity in performance of ℓ_1-regularization and NNLS for designs with a self-regularizing property. We show that the latter admits a reduction to the scheme pioneered in [5] to establish near-optimal performance guarantees of the lasso and the related Dantzig selector [13] with respect to estimation of β^* and and prediction under a sparsity scenario. Similar results are shown e.g. in [10, 12, 13, 36, 51, 58, 59], and we shall state results of that flavour for NNLS below. Throughout the rest of the paper, the data-generating model (1.1) is considered for a sparse target β^* with support $S = \{j : \beta^*_j > 0\}$, $1 \leq |S| = s < n$.

Reduction to the scheme used for the lasso. As stated in the next lemma, if the design satisfies Condition 1, the NNLS estimator $\widehat{\beta}$ has, with high probability, the property that $\widehat{\delta} = \widehat{\beta} - \beta^*$ has small ℓ_1-norm, or that $\widehat{\delta}$ is contained in the convex cone

$$\{\delta \in \mathbb{R}^p : \|\delta_{S^c}\|_1 \leq c_0 \|\delta_S\|_1\}, \quad \text{where } c_0 = \frac{3}{\tau^2}. \tag{3.1}$$

The latter property is shared by the lasso and Dantzig selector with different values of the constant c_0 [5, 13].

Lemma 1. *Assume that $y = X\beta^* + \varepsilon$, where $\beta^* \succeq 0$ has support S, ε has i.i.d. zero-mean sub-Gaussian entries with parameter σ. Further assume that X satisfies Condition 1. Denote $\widehat{\delta} = \widehat{\beta} - \beta^*$. Then, for any $M \geq 0$, at least one of the following two events occurs with probability no less than $1 - 2p^{-M^2}$:*

$$\left\{ \|\widehat{\delta}_{S^c}\|_1 \leq \frac{3}{\tau^2} \|\widehat{\delta}_S\|_1 \right\}, \quad \text{and} \quad \left\{ \|\widehat{\delta}\|_1 \leq 4(1+M)\left(1 + \frac{3}{\tau^2}\right) \sigma \sqrt{\frac{2\log p}{n}} \right\}.$$

The rightmost event is most favourable, since it immediately yields the assertion of Theorem 2 below. On the other hand, under the leftmost event, one is in position to carry over techniques used for analyzing the lasso. When combined with the following Condition 2, near-optimal rates with regard to estimation and prediction can be obtained.

Condition 2. Let $\mathcal{J}(s) = \{J \subseteq \{1, \ldots, p\} : 1 \leq |J| \leq s\}$ and for $J \in \mathcal{J}(s)$ and $\alpha \geq 1$,

$$\mathcal{R}(J, \alpha) = \{\delta \in \mathbb{R}^p : \ \|\delta_{J^c}\|_1 \leq \alpha \|\delta_J\|_1\}.$$

We say that the design satisfies the (α, s)-**restricted eigenvalue condition** if there exists a constant $\phi(\alpha, s)$ so that

$$\min_{J \in \mathcal{J}(s)} \min_{\delta \in \mathcal{R}(J,\alpha)\setminus 0} \frac{\delta^\top \Sigma \delta}{\|\delta_J\|_2^2} \geq \phi(\alpha, s) > 0. \tag{3.2}$$

Condition 2 is introduced in [5]. It is weaker than several other conditions such as those in [13, 36] employed in the same context; for a comprehensive comparison of these and related conditions, we refer to [51]. Moreover, we note that Condition 2 is satisfied with overwhelming probability if X belongs to a rather broad class of random sub-Gaussian matrices with independent rows as long as n scales as $\Omega(s \log p)$ [39].

Using Lemma 1, the next statement follows along the lines of the analysis in [5].

Theorem 2. *In addition to the conditions of Lemma 1, assume further that X satisfies the $(3/\tau^2, s)$-restricted eigenvalue condition. It then holds for any $q \in [1, 2]$ and any $M \geq 0$ that*

$$\|\widehat{\beta} - \beta^*\|_q^q \leq \frac{2^{3q-2}}{\{\phi(3/\tau^2, s)\}^q} \left(1 + \frac{3}{\tau^2}\right)^{2q} s \left((1+M)\sigma\sqrt{\frac{2\log p}{n}}\right)^q$$

$$\frac{1}{n}\|X\widehat{\beta} - X\beta^*\|_2^2 \leq \frac{8(1+M)^2\sigma^2}{\phi(3/\tau^2, s)} \left(1 + \frac{3}{\tau^2}\right)^2 \frac{s \ \log p}{n},$$

with probability no less than $1 - 2p^{-M^2}$.

Theorem 2 parallels Theorem 7.2 in [5], establishing that the lasso adapts to the underlying sparsity, since its performance attains (apart from factors logarithmic in p) what could be achieved if the support β^* were known in advance. The rates of Theorem 2 for NNLS are the same as those for the lasso, modulo (less favourable) multiplicative constants.

The required condition on the design is a combination of the self-regularizing property and the restricted eigenvalue condition. At first glance, these two conditions might appear to be contradicting each other, since the first one is not satisfied if the off-diagonal entries of Σ are too small, while for $\alpha \geq 1$, we have $\phi(\alpha, s) \leq 2(1 - \max_{j,k,\, j \neq k} \langle X_j, X_k \rangle / n)$. We resolve this apparent contradiction in Section 5 by providing designs satisfying both conditions simultaneously. The use of Condition 2 in place of more restrictive conditions like restricted isometry

properties (RIP, [13]) used earlier in the literature turns out to be crucial here, since these conditions impose much stronger constraints on the magnitude of the off-diagonals entries of Σ as discussed in detail in [37].

A result of the same spirit as Theorem 2 is shown in the recent paper [34] by Meinshausen who has independently studied the performance of NNLS for high-dimensional linear models. That paper provides an ℓ_1-bound for estimation of β^* and a fast rate bound for prediction with better constants than those in above theorem, even though the required conditions are partly more restrictive. The ingredients leading to those bounds are the self-regularizing property, which is termed 'positive eigenvalue condition' there, and the 'compatibility condition' [51, 52] which is used in place of Condition 2. We prefer the latter here, because the 'compatibility condition' is not sufficient to establish ℓ_q-bounds for estimation for $q > 1$. As distinguished from our Theorem 2, a lower bound on the minimum non-zero coefficient of β^* is additionally required in the corresponding result in [34].

4. Estimation error with respect to the ℓ_∞ norm and support recovery by thresholding

In the present section, we directly derive bounds on the ℓ_∞-estimation error of NNLS without resorting to techniques and assumptions used in the analysis of the lasso. Instead, we build on the geometry underlying the analysis of NNLS for sparse recovery in the noiseless case [17–19, 55]. In light of the stated bounds, we subsequently study the performance of a thresholded NNLS estimator with regard to support recovery.

4.1. Main components of the analysis

In the sequel, we provide the main steps towards the results stated in this section. Basic to our approach is a decomposition of the NNLS problem into two sub-problems corresponding to the support S and the off-support S^c. For this purpose, we need to introduce the following quantities. For a given support S, let Π_S and $\Pi_S^\perp$ denote the projections on the subspace spanned by $\{X_j\}_{j\in S}$ and its orthogonal complement, respectively. In the context of the linear model (1.1), we then set

$$Z = \Pi_S^\perp X_{S^c} \quad \text{and} \quad \xi = \Pi_S^\perp \varepsilon. \tag{4.1}$$

These quantities appear in the following key lemma that contains the aforementioned decomposition of the NNLS problem.

Lemma 2. *Let Z and ξ be as defined in (4.1). Consider the two non-negative least squares problems*

$$(P1): \min_{\beta^{(P1)}\geq 0} \frac{1}{n}\|\xi - Z\beta^{(P1)}\|_2^2,$$

$$(P2): \min_{\beta^{(P2)}\geq 0} \frac{1}{n}\|\Pi_S\varepsilon + X_S\beta_S^* - X_S\beta^{(P2)} - \Pi_S X_{S^c}\widehat{\beta}^{(P1)}\|_2^2$$

with minimizers $\widehat{\beta}^{(P1)}$ *of* $(P1)$ *and* $\widehat{\beta}^{(P2)}$ *of* $(P2)$*, respectively. If* $\widehat{\beta}^{(P2)} \succ 0$*, then setting* $\widehat{\beta}_S = \widehat{\beta}^{(P2)}$ *and* $\widehat{\beta}_{S^c} = \widehat{\beta}^{(P1)}$ *yields a minimizer* $\widehat{\beta}$ *of the non-negative least squares problem* (1.2)*.*

Lemma 2 is used in the proof of Theorem 3 below in the following way. We first study the off-support problem $(P1)$ separately, establishing an upper bound on the ℓ_1-norm of its minimizer $\widehat{\beta}^{(P1)}$ in dependence of the *separating hyperplane constant* introduced in the next paragraph. Substituting that bound into $(P2)$, we conclude an upper bound on $\|\beta^*_S - \widehat{\beta}^{(P2)}\|_\infty$ and in turn, by the lemma, on $\|\widehat{\beta} - \beta^*\|_\infty$. In this second step, we exploit the fact that if $\widehat{\beta}^{(P2)} \succ 0$, it equals the corresponding unconstrained least squares estimator.

Separating hyperplane constant. To establish an upper bound on $\|\widehat{\beta}^{(P1)}\|_1$, we require a positive lower bound on the following quantity to which we refer as separating hyperplane constant, which is nothing else than the constant (2.2) introduced in the context of self-regularization designs in Section 2 with respect to the matrix Z in (4.1). The term 'separating hyperplane constant' follows the geometric interpretation as margin of a hyperplane that contains the origin and that separates $\{X_j\}_{j\in S}$ from $\{X_j\}_{j\in S^c}$. Accordingly, for given S, we define

$$\tau(S) = \left\{ \max \tau : \exists w \in \mathbb{R}^n,\ \|w\|_2 \leq 1 \ \text{ s.t. } \ \frac{1}{\sqrt{n}} X_S^\top w = 0 \ \text{ and } \ \frac{1}{\sqrt{n}} X_{S^c}^\top w \succeq \tau \mathbf{1} \right\}. \tag{4.2}$$

By convex duality, we have

$$\begin{aligned} \tau^2(S) &= \min_{\substack{\theta \in \mathbb{R}^s \\ \lambda \in T^{p-s-1}}} \frac{1}{n} \|X_S\theta - X_{S^c}\lambda\|_2^2, \ \text{ where } T^{p-s-1} = \{\lambda \in \mathbb{R}_+^{p-s} : \ \lambda^\top \mathbf{1} = 1\} \\ &= \min_{\lambda \in T^{p-s-1}} \lambda^\top \frac{1}{n} X_{S^c}^\top \Pi_S^\perp X_{S^c} \lambda = \min_{\lambda \in T^{p-s-1}} \lambda^\top \frac{1}{n} Z^\top Z \lambda. \end{aligned} \tag{4.3}$$

The last line highlights the connection to (2.3) in Section 2. Expanding $\frac{1}{n} Z^\top Z$ under the assumption that the submatrix Σ_{SS} is invertible, $\tau^2(S)$ can also be written as

$$\tau^2(S) = \min_{\lambda \in T^{p-s-1}} \lambda^\top \left(\Sigma_{S^cS^c} - \Sigma_{S^cS} \Sigma_{SS}^{-1} \Sigma_{SS^c} \right) \lambda \tag{4.4}$$

It is shown in [44] that having $\tau(S) > 0$ is a necessary and sufficient condition for recovery of β^* by NNLS in the absence of noise ($\varepsilon = 0$). Thus, the appearance of $\tau(S)$ in the present context is natural.

4.2. Bounds on the ℓ_∞-error

The upper bound of the next theorem additionally depends on the quantities below, which also appear in the upper bound on the ℓ_∞-error of the lasso [54].

$$\beta_{\min}(S) = \min_{j \in S} \beta_j^*, \quad K(S) = \max_{\|v\|_\infty = 1} \|\Sigma_{SS}^{-1} v\|_\infty, \quad \phi_{\min}(S) = \min_{\|v\|_2 = 1} \|\Sigma_{SS} v\|_2 . \tag{4.5}$$

Theorem 3. *Assume that $y = X\beta^* + \varepsilon$, where $\beta^* \succeq 0$ and ε has i.i.d. zero-mean sub-Gaussian entries with parameter σ. For $M \geq 0$, set*

$$b = \frac{2(1+M)\sigma\sqrt{\frac{2\log p}{n}}}{\tau^2(S)}, \quad \text{and} \quad \widetilde{b} = b \cdot K(S) + \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(S)}}\sqrt{\frac{2\log p}{n}}. \tag{4.6}$$

If $\beta_{\min}(S) > \widetilde{b}$, then the NNLS estimator $\widehat{\beta}$ has the following properties with probability no less than $1 - 4p^{-M^2}$:

$$\|\widehat{\beta}_{S^c}\|_1 \leq b \quad \text{and} \quad \|\widehat{\beta}_S - \beta^*_S\|_\infty \leq \widetilde{b}.$$

Discussion. Theorem 3 can be summarized as follows. Given a sufficient amount of separation between $\{X_j\}_{j\in S}$ and $\{X_j\}_{j\in S^c}$ as quantified by $\tau^2(S)$, the ℓ_1-norm of the off-support coefficients is upper bounded by the effective noise level proportional to $\sqrt{\log(p)/n}$ divided by $\tau^2(S)$, provided that the entries of β^*_S are all large enough. The upper bound $\widetilde{b}$ depends in particular on the ratio $K(S)/\tau^2(S)$. In Section 5, we discuss a rather special design for which $\tau^2(S) = \Omega(1)$; for a broader class of designs that is shown to satisfy the conditions of Theorem 2 as well, $\tau^2(S)$ roughly scales as $\Omega(s^{-1})$. Moreover, we have $\{\phi_{\min}(S)\}^{-1} \leq K(S) \leq \sqrt{s}\{\phi_{\min}(S)\}^{-1}$. In total, the ℓ_∞-bound can hence be as large as $O(s^{3/2}\sqrt{\log(p)/n})$ even if $\tau^2(S)$ scales favourably, a bound that may already be implied by the ℓ_2-bound in Theorem 2. On the positive side, Theorem 3 may yield a satisfactory result for s constant or growing only slowly with n, without requiring the restricted eigenvalue condition of Theorem 2.

Towards a possible improvement of Theorem 3. The potentially suboptimal dependence on the sparsity level s in the bounds of Theorem 3 is too pessimistic relative to the empirical behaviour of NNLS as discussed in Section 6. The performance reported there can be better understood in light of Theorem 4 below and the comments that follow. Our reasoning is based on the fact that any NNLS solution can be obtained from an ordinary least squares solution restricted to the variables in the active set $F = \{j : \widehat{\beta}_j > 0\}$, cf. Lemma 3 in Appendix E. For the subsequent discussion to be meaningful, it is necessary that the NNLS solution and thus its active set are unique, for which a sufficient condition is thus established along the way.

Theorem 4. *Let the data-generating model be as in Theorem 3 and let $M \geq 0$ be arbitrary. If the columns of X are in general linear position (1.5) and if*

$$\frac{32(1+M)^2\sigma^2}{\mathbf{E}[\varepsilon_1^2]} \; \frac{\log p}{\tau^2(S)\, n} \leq \left(1 - \frac{s}{n}\right), \tag{4.7}$$

then, with probability at least $1 - \exp(-c(n-s)/\sigma^4) - 2p^{-M^2}$, the NNLS solution is unique and its active set $F = \{j : \widehat{\beta}_j > 0\}$ satisfies $|F| \leq \min\{n-1, p\}$. Conditional on that event, if furthermore $\beta_{\min}(S) > \widetilde{b}$ as defined in (4.6), then

$S \subseteq F$ and

$$\|\widehat{\beta} - \beta^*\|_\infty \leq \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(F)}}\sqrt{\frac{2\log p}{n}}, \tag{4.8}$$

with probability at least $1 - 6p^{-M^2}$.

We first note that for s/n bounded away from 1, condition (4.7) is fulfilled if n scales as $\Omega(\log(p)/\tau^2(S))$. Second, the condition on $\beta_{\min}(S)$ is the same as in the previous Theorem 3, so that the scope of application of the above theorem remains limited to designs with an appropriate lower bound on $\tau^2(S)$. At the same time, Theorem 4 may yield a significantly improved bound on $\|\widehat{\beta} - \beta^*\|_\infty$ as compared to Theorem 3 if $\{\phi_{\min}(F)\}^{-1/2}$, the smallest singular value of $X_F/\sqrt{n} \in \mathbb{R}^{n\times|F|}$, scales more favourably than $K(S)/\tau^2(S)$, noting that as long as $S \subseteq F$, $\{\phi_{\min}(S)\}^{-1/2} \leq \{\phi_{\min}(F)\}^{-1/2}$. In the first place, control of $\{\phi_{\min}(F)\}^{-1/2}$ requires control over the cardinality of the set F. In a regime with $|F|$ scaling as a constant multiple of s with $s = \alpha n$, $0 \leq \alpha \ll 1$, it is not restrictive to assume that $\{\phi_{\min}(F)\}^{1/2}$ as the smallest singular value of a tall submatrix of X is lower bounded by a positive constant, as it has been done in the literature on ℓ_1-regularization [13, 36, 58]. That assumption is strongly supported by results in random matrix theory [32, 38]. In Section 5 the hypothesis of having $|F| \ll n$ is discussed in more detail for the class of so-called equi-correlation-like designs. For equi-correlated design, it is even possible to derive the distribution of $|F|$ conditional on having $S \subseteq F$ (Proposition 2 in Section 5).

4.3. Support recovery by thresholding

The bounds on the estimation error presented in the preceding two sections imply that hard thresholding of the NNLS estimator may be an effective means for recovery of the support S. Formally, for a threshold $t \geq 0$, the hard-thresholded NNLS estimator is defined by

$$\widehat{\beta}_j(t) = \begin{cases} \widehat{\beta}_j, & \widehat{\beta}_j > t, \\ 0, & \text{otherwise}, \ j = 1, \ldots, p, \end{cases} \tag{4.9}$$

and we consider $\widehat{S}(t) = \{j : \widehat{\beta}_j > 0\}$ as an estimator for S. In principle, the threshold might be chosen according to Theorem 3 or 4: if $t > b$ and $\beta_{\min}(S) > b + \widetilde{b}$, where b and $\widetilde{b}$ denote upper bounds on $\|\widehat{\beta}_{S^c}\|_\infty$ and $\|\widehat{\beta}_S - \beta^*_S\|_\infty$, respectively, one has that $S = \widehat{S}(t)$ with the stated probabilities. This approach, however, is not practical, since the bounds b and $\widetilde{b}$ depend on constants that are not accessible. In the sequel, we propose a data-driven approach as devised in [23] for support recovery on the basis of marginal regression. A central observation in [23] is that direct specification of the threshold can be avoided if the purpose of thresholding is support recovery. In fact, given a ranking $(r_j)_{j=1}^p$ of the predictors $\{X_j\}_{j=1}^p$ so that $r_j \leq s$ for all $j \in S$, it suffices to estimate s. In

light of Theorems 2 to 4, NNLS may give rise to such ranking by setting

$$r_j = k \iff \widehat{\beta}_j = \widehat{\beta}_{(k)},\ j = 1, \ldots, p, \tag{4.10}$$

where $\widehat{\beta}_{(1)} \geq \widehat{\beta}_{(2)} \geq \cdots \geq \widehat{\beta}_{(p)}$ is the sequence of coefficients arranged in decreasing order. Theorem 5 below asserts that conditional on having an ordering in which the first s variables are those in S, support recovery can be achieved by using the procedure in [23]. Unlike the corresponding result in [23], our statement is non-asymptotic and comes with a condition that is easier to verify. We point out that Theorem 5 is of independent interest, since it is actually not specific to NNLS, but would equally apply to any estimator yielding the correct ordering of the variables.

Theorem 5. *Consider the data-generating model of Theorem 3 and suppose that the NNLS estimator has the property that according to (4.10), it holds that $r_j \leq s$ for all $j \in S$. For any $M \geq 0$, set*

$$\begin{aligned} \widehat{s} &= \max\left\{0 \leq k \leq (p-1) : \delta(k) \geq (1+M)\sigma\sqrt{2\ \log p}\right\} + 1, \\ \text{where}\ \ \delta(k) &= \left\|(\Pi(k+1) - \Pi(k))y\right\|_2,\ k = 0, \ldots, (p-1), \end{aligned} \tag{4.11}$$

with $\Pi(k)$ denoting the projection on the linear space spanned by the variables whose ranks are no larger than k (using $\Pi(0) = 0$). Let $\widehat{S} = \{j : r_j \leq \widehat{s}\}$.

If $\beta_{\min}(S) \geq 2(1+M)\sigma\{\phi_{\min}(S)\}^{-1/2}\sqrt{2\log(p)/n}$, then $\widehat{S} = S$ with probability no less than $1 - 4p^{-M^2}$.

We note that the required lower bound on $\beta_{\min}(S)$ is rather moderate. Similar or even more stringent lower bounds are required throughout the literature on support recovery in a noisy setup [9, 12, 33, 54, 59, 60], and are typically already needed to ensure that the variables in S are ranked at the top (cf. also Theorems 2 to 4).

Strictly speaking, the estimate $\widehat{s}$ in Theorem 5 is not operational, since knowledge of the noise level σ is assumed. In practice, σ has to be replaced by a suitable estimator. Variance estimation in high-dimensional linear regression with Gaussian errors continues to be a topic of active research, with several significant advances made very recently [26]. In our experiments, this issue appears to be minor, because even naive plug-in estimation of the form $\widehat{\sigma}^2 = \frac{1}{n}\|y - X\widehat{\beta}\|_2^2$ yields satisfactory results [2](cf. Section 6). A nice property of the approach is its computational simplicity. Repeated evaluation of $\delta(k)$ in (4.11) can be implemented efficiently by updating QR decompositions. Finally, we note that subsequent to thresholding, it is beneficial to re-compute the NNLS solution using data $(y, X_{\widehat{S}})$ only, because the removal of superfluous variables leads to a more accurate estimation of the support coefficients.

[2]We note that the denominator n could be replaced by $n - \nu$, with ν denoting the degrees of freedom of NNLS (which, to the best of our knowledge, is not known).

4.4. Comparison of NNLS and the non-negative lasso

Let us recall the non-negative lasso problem (1.3) given by

$$\min_{\beta \geq 0} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \mathbf{1}^\top \beta, \quad \lambda > 0,$$

with minimizer denoted by $\widehat{\beta}^{\ell_1,\geq}$. In the present subsection, we elaborate on similarities and differences of NNLS and the non-negative lasso regarding the ℓ_∞-error in estimating β^*. We first state a result according to which the non-negative lasso succeeds in support recovery. We then argue that in general, the non-negative lasso does not attain the optimal rate $O(\sqrt{\log(p)/n})$ in estimation with respect to the ℓ_∞-norm by providing a specific design as counterexample. We finally conclude by summarizing benefits and drawbacks of the non-negative lasso and NNLS in an informal way.

Non-negative irrepresentable condition. We study the performance of the non-negative lasso estimator $\widehat{\beta}^{\ell_1,\geq}$ under a version of the 'irrepresentable condition' that takes into account non-negativity of the regression coefficients. The irrepresentable condition has been employed e.g. in [35, 52, 54, 61] to study the (ordinary) lasso from the point of view of support recovery. For given $S \subset \{1, \ldots, p\}$, the *non-negative irrepresentable constant* is defined as

$$\iota(S) = \max_{j \in S^c} \Sigma_{jS} \Sigma_{SS}^{-1} \mathbf{1} = \max_{j \in S^c} X_j^\top X_S (X_S^\top X_S)^{-1} \mathbf{1}. \tag{4.12}$$

It follows from the analysis in [54] that the non-negative irrepresentable condition $\iota(S) < 1$ is necessary for the non-negative lasso to recover the support S, cf. Theorem 6 below.

Remark 2. We here point out that the condition $\iota(S) < 1$ can be regarded as a strengthening of the condition

$$\exists w \in \mathbb{R}^n \text{ s.t. } X_S^\top w = \mathbf{1} \text{ and } X_{S^c}^\top w \prec \mathbf{1}, \tag{4.13}$$

which is a necessary condition for recovering a non-negative solution with support S as minimum ℓ_1-norm solution of an underdetermined linear system of equations [62], which corresponds to a non-negative lasso problem in the absence of noise. The non-negative irrepresentable condition $\iota(S) < 1$ results from (4.13) by additionally requiring the vector w to be contained in the column space of X_S. Condition (4.13) highlights a conceptual connection to the separating hyperplane constant as defined in (4.2). Note that as distinguished from (4.2), the separating hyperplane underlying (4.13) does **not** contain the origin.

Support recovery with the non-negative lasso. Using the scheme developed in [54], one can show that under the non-negative irrepresentable condition and a suitable choice of the regularization parameter λ, the non-negative lasso has the property that $\widehat{\beta}_{S^c}^{\ell_1,\geq} = 0$, i.e. no false positive variables are selected, and

support recovery can be deduced from having an appropriate lower bound on the minimum support coefficient $\beta_{\min}(S)$. Along with a bound on $\|\widehat{\beta}^{\ell_{1,\succeq}} - \beta^*\|_\infty$, this is stated in the next theorem.

Theorem 6. *Assume that $y = X\beta^* + \varepsilon$, where $\beta^* \succeq 0$ has support S and ε has i.i.d. sub-Gaussian entries with parameter σ. Suppose further that the non-negative irrepresentable condition $\iota(S) < 1$ according to (4.12) holds. For any $M \geq 0$, if*

$$\begin{aligned} &\lambda > \frac{2\lambda_M}{1-\iota(S)}, \quad \textit{where} \quad \lambda_M = (1+M)\,\sigma\sqrt{\frac{2\log p}{n}}, \\ &\textit{and} \quad \beta_{\min}(S) > b, \quad \textit{where} \quad b = \frac{\lambda}{2}\|\Sigma_{SS}^{-1}\mathbf{1}\|_\infty + \frac{\lambda_M}{\sqrt{\phi_{\min}(S)}}, \end{aligned} \tag{4.14}$$

then $\{j : \widehat{\beta}_j^{\ell_{1,\succeq}} > 0\} = S$ and $\|\widehat{\beta}_S^{\ell_{1,\succeq}} - \beta_S^*\|_\infty \leq b$ with probability at least $1 - 4p^{-M^2}$.

There is some resemblance of the bound b in (4.14) and that of Theorem 3 for NNLS, with $\tau^2(S)$ playing a role comparable to $1 - \iota(S)$ and $\|\Sigma_{SS}^{-1}\mathbf{1}\|_\infty$ being a lower bound on the quantity $K(S)$ defined in (4.5). On the other hand, Theorem 6 yields a considerably stronger control of the off-support coefficients ($\widehat{\beta}_{S^c}^{\ell_{1,\succeq}} = 0$) as does Theorem 3, which only provides an ℓ_1-bound on $\widehat{\beta}_{S^c}$. Irrepresentable conditions as in above theorem are regarded as rather restrictive in the literature [36, 58, 60]. Even in case the condition $\iota(S) < 1$ is fulfilled, the choice of λ in (4.14) with $\iota(S)$ possibly close to one may impose a rather stringent lower bound on $\beta_{\min}(S)$ in order to achieve support recovery. At the same time, the choice $\lambda = 2\sigma\sqrt{2\log(p)/n}$ in combination with the restricted eigenvalue condition (Condition 2), which is regarded as far less restrictive than the irrepresentable condition, only yields a bound on $\|\widehat{\beta}^{\ell_{1,\succeq}} - \beta^*\|_q$ for $q \in [1,2]$, and it is no longer guaranteed that $\widehat{\beta}_{S^c}^{\ell_{1,\succeq}} = 0$. As a result, two-stage procedures like subsequent thresholding of $\widehat{\beta}^{\ell_{1,\succeq}}$ may be needed for support recovery. However, this approach in general entails a sub-optimal condition on $\beta_{\min}(S) = \Omega(\sqrt{s\log(p)/n})$ because of the term $\|\Sigma_{SS}^{-1}\mathbf{1}\|_\infty$ scaling as $\Theta(\sqrt{s})$ in the worst case. In Appendix J this issue is illustrated by providing an explicit example of a design representing that worst case. We point out that optimal rates for estimation in sup-norm of the order $O(\sqrt{\log(p)/n})$ have been established e.g. in [9, 12, 33] for the lasso under 'mutual incoherence' conditions requiring fairly restrictive upper bounds on the maximum inner product between two distinct columns of X.

NNLS vs. the non-negative lasso: pros and cons. To sum up, we list advantages and disadvantages of NNLS and the non-negative lasso, thereby providing some guidance on when to use which approach in practice. While NNLS can formally be seen as a special case of the non-negative lasso with $\lambda = 0$, we suppose for the subsequent discussion that $\lambda \geq \sigma\sqrt{2\log(p)/n}$ as it is standard in the literature on the lasso.

- As already stressed previously, reasonable performance of NNLS in a high-dimensional regime is restricted to a specific class of designs, which excludes standard models such as random Gaussian design matrices (cf. the discussion in Section 5). This contrasts with the non-negative lasso, which has at least moderate performance guarantees via a slow rate bound in spirit of Theorem 1 *without* further conditions on the design.
- As discussed in the previous paragraph, the non-negative lasso does not always attain the optimal rate for estimating β^* in the sup-norm, in which case some room for improvement is left for NNLS. In Section 6, we present two designs for which NNLS empirically yields a better performance both with regard to estimation and support recovery via thresholding, where the ℓ_∞-error in estimation enters crucially. On the other hand, as asserted by Theorem 6 the non-negative lasso succeeds in support recovery even without thresholding in certain regimes.
- From the point of view of a practitioner who is little familiar with theoretical results on how to set the regularization parameter of the (non-negative) lasso, NNLS has the advantage that it can be applied directly, without the need to specify or tune a regularization parameter.

5. Discussion of the analysis of NNLS for selected designs

Our main results concerning the performance of NNLS as stated in Theorems 1 to 4 are subject to the following conditions: the self-regularizing property (Theorem 1), a combination of that property with a restricted eigenvalue condition (Theorem 2), a lower bound on the separating hyperplane constant (Theorem 3), and sparsity of the NNLS solution (Theorem 4). In the present section, we discuss to what extent these conditions are fulfilled for selected designs, which we here roughly divide into three classes. The first is the class of *non-self-regularizing designs* for which non-negativity constraints on the regression coefficients do not seem to yield any significant advantage. This is in contrast to the third class of *equi-correlation-like designs*, which are shown to be tailored to NNLS. The second class comprises designs with a block or band structure arising in typical applications.

5.1. Non-self regularizing designs

In this paragraph, we provide several common examples of designs not having the self-regularizing property of Condition 1. Consequently, our main results, which rely on that condition, do not apply. Those designs can be identified by evaluating the quantity τ_0^2 (2.3) underlying Condition 1. From

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda \leq \frac{1}{p^2} \mathbf{1}^\top \Sigma \mathbf{1}, \tag{5.1}$$

we see that the sum of the entries of Σ must scale as $\Omega(p^2)$ for Condition 1 to be satisfied. In particular, this requires Σ to have $\Omega(p^2)$ entries lower bounded by

a positive constant, and a maximum eigenvalue scaling as $\Omega(p)$. Among others, this is not fulfilled for the following examples.

Example 1 (**orthonormal design**)**.** As already mentioned while motivating the self-regularizing property in Section 2, for $\Sigma = I$, τ_0^2 attains the upper bound in (5.1) which yields $\tau_0^2 = 1/p$.

Example 2 (**power decay**)**.** Let the entries of Σ be given by $\sigma_{jk} = \rho^{|j-k|}$, $j, k = 1, \ldots, p$ with $\rho \in [0, 1)$. From

$$\max_{1 \le j \le p} \sum_{k=1}^{p} \sigma_{jk} \le \sum_{l=0}^{p-1} \rho^l \le (1-\rho)^{-1}$$

and (5.1) it follows that $\tau_0^2 \le p^{-1}(1-\rho)^{-1}$.

Example 3 (**random Gaussian matrices**)**.** Consider a random matrix X whose entries are drawn i.i.d. from the standard Gaussian distribution. We here refer to the work [19] in which NNLS is studied in the noiseless case. In that work, random Gaussian matrices are considered as part of a broader class termed the centro-symmetric ensemble. In a nutshell, that class encompasses random matrices with independent columns whose entries have mean zero. The authors of [19] point out the importance of Wendel's Theorem [40, 57], which provides the exact probability for the columns of X being contained in a half-space, i.e. of having $\tau_0^2 > 0$. That result implies via Hoeffding's inequality that for $p/n > 2$, $\tau_0^2 = 0$ with probability at least $1 - \exp(-n(p/n-2)^2/2)$ so that the non-negativity constraints in NNLS become meaningless (cf. the discussion following (2.1)).

In all these three examples, a similar reasoning applies with regard to the scaling of the separating hyperplane constant $\tau^2(S)$ (4.3), because its role is that of τ_0^2 with respect to the matrix $Z = \Pi_S^{\perp} X_{S^c}$. As a consequence, it is not hard to see that the scalings of the above examples continue to hold (uniformly in S) with p replaced by $p - s$.

5.2. Designs with non-negative Gram matrices having a band or block structure

We now present a simple sufficient condition for the self-regularizing property to be satisfied, based on which we will identify a class of designs for which non-negativity of the regressions coefficients may be a powerful constraint.

Suppose that the Gram matrix has the property that all its entries are lower bounded by a positive constant σ_0. We then have the following lower bound corresponding to the upper bound (5.1) above.

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda \ge \min_{\lambda \in T^{p-1}} \lambda^\top \left\{ \sigma_0 \mathbf{1}\mathbf{1}^\top \right\} \lambda = \sigma_0, \tag{5.2}$$

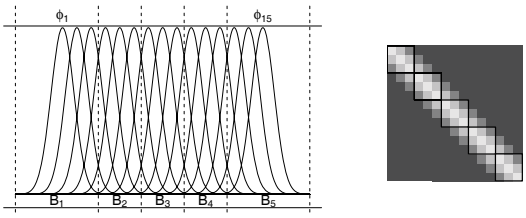

FIG 1. *Block partitioning of 15 Gaussians into $K = 5$ blocks. The right part shows the corresponding pattern of the Gram matrix.*

i.e. Condition 1 is satisfied with $\tau^2 = \sigma_0$. More generally, in case that Σ has exclusively non-negative entries and the set of variables $\{1, \ldots, p\}$ can be partitioned into blocks $\{B_1, \ldots, B_K\}$ such that the minimum entries of the corresponding principal submatrices of Σ are lower bounded by a positive constant, then Condition 1 is satisfied with $\tau^2 = \sigma_0/K$:

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda \geq \min_{\lambda \in T^{p-1}} \sum_{l=1}^K \lambda_{B_l}^\top \Sigma_{B_l\, B_l} \lambda_{B_l} \geq \sigma_0 \min_{\lambda \in T^{p-1}} \sum_{l=1}^K (\lambda_{B_l}^\top \mathbf{1})^2 = \sigma_0/K, \tag{5.3}$$

where in the last equality we have used that the minimum of the map $x \mapsto \sum_{l=1}^K x_l^2$ over the simplex T^{K-1} is attained for $x = \mathbf{1}/K$.

As sketched in Figure 1, the lower bound (5.3) particularly applies to design matrices whose entries contain the function evaluations at points $\{u_i\}_{i=1}^n \subset [a, b]$ of non-negative functions such as splines, Gaussian kernels and related 'localized' functions traditionally used for data smoothing. If the points $\{u_i\}_{i=1}^n$ are placed evenly in $[a, b]$ then the corresponding Gram matrix effectively has a band structure. For instance, suppose that $u_i = i/n$, $i = 1, \ldots, n$, and consider indicator functions of sub-intervals $\phi_j(u) = I\{u \in [(\mu_j - h) \vee a, (\mu_j + h) \wedge b]\}$, where $\mu_j \in [0, 1]$, $j = 1, \ldots, p$, and $h = 1/K$ for some positive integer K. Setting $X = (\phi_j(u_i))_{1 \leq i \leq n,\, 1 \leq j \leq p}$ and partitioning the $\{\mu_j\}$ by dividing $[0, 1]$ into intervals $[0, h]$, $(h, 2h]$, $\ldots$, $(1 - h, 1]$ and accordingly $B_l = \{j : \mu_j \in ((l - 1) \cdot h, l \cdot h]\}$, $l = 1, \ldots, K$, we have that $\min_{1 \leq l \leq K} \frac{1}{n} X_{B_l}^\top X_{B_l} \succeq h$ such that Condition 1 holds with $\tau^2 = h/K = 1/K^2$.

Applications. As mentioned in the introduction, NNLS has been shown to be remarkably effective in solving deconvolution problems [30, 31, 42]. The observations there are signal intensities measured over time, location etc. that can be modelled as a series of spikes (Dirac impulses) convolved with a *point-spread function (PSF)* arising from a limited resolution of the measurement device. The PSF is a non-negative localized function as outlined in the previous paragraph. Deconvolution of spike trains is studied in more detail in Section 6.1 below. Similarly, bivariate PSFs can be used to model blurring in greyscale images, and NNLS has been considered as a simple method for deblurring and denoising [2].

5.3. Equi-correlation-like designs

We first discuss equi-correlated design before studying random designs whose population Gram matrix has equi-correlation structure. While the population setting is limited to having $n \geq p$, the case $n < p$ is possible for random designs.

Equi-correlated design. For $\rho \in (0,1)$, consider equi-correlated design with Gram matrix $\Sigma = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$. We then have

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda = \rho + \min_{\lambda \in T^{p-1}} (1-\rho) \|\lambda\|_2^2 = \rho + \frac{1-\rho}{p}, \tag{5.4}$$

so that the design has the self-regularizing property of Condition 1. Let $\emptyset \neq S \subset \{1,\dots,p\}$ be arbitrary. According to representation (4.4), the corresponding separating hyperplane constant $\tau^2(S)$ can be evaluated similarly to (5.4). We have

$$\begin{aligned} \tau^2(S) &= \min_{\lambda \in T^{p-s-1}} \lambda^\top \left(\Sigma_{S^cS^c} - \Sigma_{S^cS} \Sigma_{SS}^{-1} \Sigma_{SS^c} \right) \lambda \\ &= \rho - \rho^2 \mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{1} + (1-\rho) \min_{\lambda \in T^{p-s-1}} \|\lambda\|_2^2 \\ &= \rho - \frac{s\rho^2}{1+(s-1)\rho} + \frac{1-\rho}{p-s} = \frac{\rho(1-\rho)}{1+(s-1)\rho} + \frac{1-\rho}{p-s} = \Omega(s^{-1}), \end{aligned} \tag{5.5}$$

where from the second to the third line we have used that $\mathbf{1}$ is an eigenvector of Σ_{SS} corresponding to its largest eigenvalue $1+(s-1)\rho$. We observe that $\tau^2(S) = \tau^2(s)$, i.e. (5.5) holds uniformly in S. We are not aware of any design for which $\min_{S:|S|=s<p/2} \tau^2(S) \geq s^{-1}$, which lets us hypothesize that the scaling of $\tau^2(S)$ in (5.5) uniformly over all sets of a fixed cardinality s is optimal. On the other hand, when not requiring uniformity in S, $\tau^2(S)$ can be as large as a constant independent of s, as it is the case for the following example. Consider a Gram matrix of the form

$$\Sigma = \begin{pmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^cS} & \Sigma_{S^cS^c} \end{pmatrix} = \begin{pmatrix} \Sigma_{SS} & \mathbf{0} \\ \mathbf{0} & (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top \end{pmatrix} \quad \text{for } \rho \in (0,1).$$

Combining (5.4) and (5.5), we obtain that $\tau^2(S) = \rho + \frac{1-\rho}{p-s}$ independently of the specific form of Σ_{SS}. At the same time, this scaling does not hold uniformly over all choices of S with $|S| = s$ given the equi-correlation structure of the block $\Sigma_{S^cS^c}$.

Sparsity of the NNLS solution for equi-correlated design. Exploiting the specifically simple structure of the Gram matrix, we are able to derive the distribution of the cardinality of the active set $F = \{j : \widehat{\beta}_j > 0\}$ of the NNLS solution $\widehat{\beta}$ conditional on the event $\{\widehat{\beta}_S \succ 0\}$. For the sake of better illustration, the result is stated under the assumption of Gaussian noise. Inspection of the proof shows that, with appropriate modifications, the result remains valid for arbitrary noise distributions.

Proposition 2. *Consider the linear model* $y = X\beta^* + \varepsilon$, *where* $\beta^* \succeq 0$, $\frac{1}{n}X^\top X = \Sigma = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top$ *for* $\rho \in [0,1)$, *and* ε *has i.i.d. zero-mean, Gaussian entries with variance* σ^2. *Let further* $S = \{j : \beta_j^* > 0\}$. *For any* $M \geq 0$, *if* $\beta_{\min}(S) > \frac{3(1+M)\sigma}{1-\rho}\sqrt{2\log(p)/n}$, *then the event* $\{\widehat{\beta}_S \succ 0\}$ *occurs with probability at least* $1 - 4p^{-M^2}$. *Furthermore, let* z *be a* $(p-s)$*-dimensional zero-mean Gaussian random vector with covariance* $(1-\rho)I + \frac{\rho(1-\rho)}{1+(s-1)\rho}\mathbf{1}\mathbf{1}^\top$ *and let* $z_{(1)} \geq \cdots \geq z_{(p-s)}$ *denote the arrangement of the components of* z *in decreasing order. Conditional on the event* $\{\widehat{\beta}_S \succ 0\}$, *the cardinality of the active set* $F = \{j : \widehat{\beta}_j > 0\}$ *has the following distribution:*

$$|F| \stackrel{\mathcal{D}}{=} s + I\left\{z_{(1)} > 0\right\}\left(1 + \max\left\{1 \leq j \leq p-s-1 : \zeta_j > \theta(s,\rho)\right\}\right), \text{ where}$$
$$\zeta_j = \frac{z_{(j+1)}}{\sum_{k=1}^{j}(z_{(k)} - z_{(j+1)})}, \quad j = 1,\ldots,p-s-1, \quad \text{and } \theta(s,\rho) = \frac{\rho}{1+(s-1)\rho}. \tag{5.6}$$

Proposition 2 asserts that conditional on having the support of β^* included in the active set, the distribution of its cardinality is s plus an extra term, whose distribution depends on that of the random variables $\{\zeta\}_{j=1}^{p-s-1}$ and a 'threshold' $\theta(s,\rho)$. In order to better understand the role of these quantities, let us first consider the case $\rho = 0$, i.e. orthonormal design: since $\theta(s,0) = 0$, the distribution of $|F|$ is equal to s plus the distribution of the number of non-negative components of a $(p-s)$-dimensional Gaussian random vector, i.e. a binomial distribution with $p-s$ trials and a probability of success of $\frac{1}{2}$ (cf. also Section 2.2). Once $\rho > 0$, the distribution of $|F|$ gets shifted towards s, noting that $\{\zeta\}_{j=1}^{p-s-1}$ forms a non-increasing sequence. Specifically, for $s = 0$, $\theta(0,\rho) = \frac{\rho}{1-\rho}$, i.e. the larger the correlation ρ, the stronger the concentration of the distribution of $|F|$ near zero. The threshold $\theta(s,\rho)$ is decreasing in s, i.e. the number of extra variables increases with s. While the distribution of $\{\zeta\}_{j=1}^{p-s-1}$ is not directly accessible, it can be approximated arbitrarily well by Monte Carlo simulation for given p, s and ρ (note that the distribution does not depend on the scale of the noise ε). Figure 2 depicts the $0.01, 0.5, 0.99$-quantiles of the distribution of $|F|$ in virtue of (5.6) for $p = 500$ and various choices of ρ and s. The results are based on $10,000$ Monte Carlo simulations for each value of s. For comparison, for each pair (s,ρ), we generate 100 datasets (X,y) with $n = p = 500$ according to the model of Proposition 2 with standard Gaussian noise (the components of β_S^* are set to the given lower bound on $\beta_{\min}(S)$ in to ensure that the event $\{\widehat{\beta}_S \succ 0\}$ has probability close to one). We then solve the corresponding NNLS problems using the active set algorithm of Lawson and Hanson [29] and obtain the cardinalities of the active sets. Figure 2 shows a strong agreement of the predictions regarding the size of the active set based on the distribution of Proposition 2 and the empirical distributions.

Non-negative random designs with equi-correlation structure. We now consider random design matrices whose population Gram matrix is that

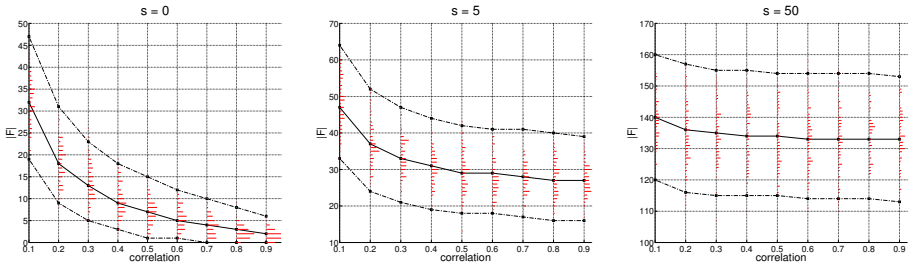

FIG 2. *Graphical illustration of Proposition 2 for* $p = 500$. *The dotted lines represent the* $\{0.01, 0.5, 0.99\}$*-quantiles of the distributions obtained from Proposition 2 via Monte Carlo simulation. The horizontal bars represent the corresponding relative frequencies based on the solutions of* 100 *random NNLS problems obtained for each combination of* ρ *and* s.

of equi-correlated design, but with the possibility that $n < p$. It is investigated to what extent these random design matrices inherit properties from the population setting studied in the previous paragraph. Specifically, we consider the following ensemble of random matrices

$$\mathrm{Ens}_+ : \; X = (x_{ij})_{\substack{1 \le i \le n, \\ 1 \le j \le p}}, \; \{x_{ij}\} \text{ i.i.d. from a sub-Gaussian distribution on } \mathbb{R}_+. \tag{5.7}$$

All random designs from the class Ens_+ share the property that the population Gram matrix $\Sigma^* = \mathbf{E}[\frac{1}{n} X^\top X]$ possesses equi-correlation structure after re-scaling the entries of X by a common factor. Denoting the mean of the entries and their squares by μ and μ_2, respectively, we have

$$\Sigma^* = \mathbf{E}\left[\frac{1}{n} X^\top X\right] = (\mu_2 - \mu^2) I + \mu^2 \mathbf{1}\mathbf{1}^\top,$$

such that re-scaling by $1/\sqrt{\mu_2}$ leads to equi-correlation structure with parameter $\rho = \mu^2/\mu_2$. Since applications of NNLS predominantly involve non-negative design matrices, it is instructive to have a closer look at the class (5.7) as a basic model for such designs. Among others, the class of sub-Gaussian random designs on $\mathbb{R}_+$ encompasses the zero-truncated Gaussian distribution, all distributions on a bounded subset of $\mathbb{R}_+$, e.g. the family of beta distributions (with the uniform distribution as special case) on $[0, 1]$, Bernoulli distributions on $\{0, 1\}$ or more generally multinomial distributions on positive integers $\{0, 1, \ldots, K\}$, as well as any finite mixture of these distributions.

As shown in the sequel, the class (5.7) provides instances of designs for which Theorems 2 to 4 yield meaningful results in the $n < p$ setting. Our reasoning hinges on both theoretical analysis providing bounds on the deviation from population counterparts as well as on numerical results.

Self-regularizing property + restricted eigenvalue condition of Theorem 2. Recall that Theorem 2 requires a combination of the self-regularizing property (Condition 1) and the restricted eigenvalue condition (Condition 2) to be satisfied. This turns out to be the case for designs from Ens_+ in light of he following

proposition. The statement relies on recent work of Rudelson and Zhou [39] on the restricted eigenvalue condition for random matrices with independent sub-Gaussian rows.

Proposition 3. *Let X be a random matrix from Ens_+ (5.7) scaled such that $\Sigma^* = \mathbf{E}[\frac{1}{n}X^\top X] = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top$ for some $\rho \in (0,1)$. Set $\delta \in (0,1)$. There exists constants $C, c > 0$ depending only on δ, ρ and the sub-Gaussian parameter of the centered entries of X so that if $n \geq C\, s\log(p \vee n)$, then, with probability at least $1 - \exp(-c\delta^2 n) - 6/(p \vee n)$, $\Sigma = X^\top X/n$ has the self-regularizing property with $\tau^2 = \rho/2$ and satisfies the $(3/\tau^2, s)$ restricted eigenvalue condition of Theorem 2 with $\phi(3/\tau^2, s) = (1-\rho)(1-\delta)^2$.*

Scaling of $\tau^2(S)$. The next proposition controls the deviation of the separating hyperplane constant $\tau^2(S)$ from its population counterpart as derived in (5.5).

Proposition 4. *Let X be a random matrix from Ens_+ (5.7) scaled such that $\Sigma^* = \mathbf{E}[\frac{1}{n}X^\top X] = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top$ for some $\rho \in (0,1)$. Fix $S \subset \{1,\ldots,p\}$, $|S| = s$. Then there exists constants $c, c', C, C' > 0$ depending only on ρ and the sub-Gaussian parameter of the centered entries of X such that for all $n \geq Cs^2\log(p \vee n)$,*

$$\tau^2(S) \geq cs^{-1} - C'\sqrt{\frac{\log p}{n}}$$

with probability no less than $1 - 6/(p \vee n) - 3\exp(-c'(s \vee \log n))$.

It turns out that the requirement on the sample size as indicated by Proposition 4 is too strict in light of the results of complementary numerical experiments. For these experiments, $n = 500$ is kept fixed and $p \in (1.2, 1.5, 2, 3, 5, 10)\cdot n$ and $s \in (0.01, 0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5)\cdot n$ vary. For each combination of (p,s) and several representatives of Ens_+ (re-scaled such that the population Gram matrix has equi-correlation structure), 100 random design matrices are generated. We set $S = \{1,\ldots,s\}$, compute $Z = (I - \Pi_S)X_{S^c}$ using a QR decomposition of X_S and then solve the quadratic program $\min_{\lambda \in T^{p-s-1}} \lambda^\top \frac{1}{n} Z^\top Z\lambda$ with value $\tau^2(S)$ by means of an interior point method [6]. As representatives of Ens_+, we have considered matrices whose entries have been drawn from the following distributions. In order to obtain population Gram matrices of varying correlation ρ, we use mixture distributions with one of two mixture components being a point mass at zero (denoted by δ_0). Note that the larger the proportion $1-a$ of that component, the smaller ρ.

E_1: $\{x_{ij}\} \overset{\text{i.i.d.}}{\sim} a\,\mathrm{uniform}([0, 1/\sqrt{3\cdot a}]) + (1-a)\delta_0$, $a \in \{1, \frac{2}{3}, \frac{1}{3}, \frac{2}{15}\}$ $(\rho \in \{\frac{3}{4}, \frac{1}{2}, \frac{1}{3}, \frac{1}{10}\})$

E_2: $\{x_{ij}\} \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{\pi}}\,\mathrm{Bernoulli}(\pi)$, $\pi \in \{\frac{1}{10}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{9}{10}\}$ $(\rho \in \{\frac{1}{10}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{9}{10}\})$

E_3: $\{x_{ij}\} \overset{\text{i.i.d.}}{\sim} |Z|$, $Z \sim a\,\mathrm{Gaussian}(0,1) + (1-a)\delta_0$, $a \in \{1, \frac{\pi}{4}, \frac{\pi}{8}, \frac{\pi}{20}\}$ $(\rho \in \{\frac{2}{\pi}, \frac{1}{2}, \frac{1}{4}, \frac{1}{10}\})$

E_4: $\{x_{ij}\} \overset{\text{i.i.d.}}{\sim} a\mathrm{Poisson}(3/\sqrt{12\,a}) + (1-a)\delta_0$, $a \in \{1, \frac{2}{3}, \frac{1}{3}, \frac{2}{15}\}$ $(\rho \in \{\frac{3}{4}, \frac{1}{2}, \frac{1}{4}, \frac{1}{10}\})$

For space reasons, we here only report the results for E_1. Regarding E_2 to E_4, the reader is referred to the supplementary file [45]; in brief, the results confirm what is shown here. Figure 3 displays the 0.05-quantiles of $\tau^2(S)$ over

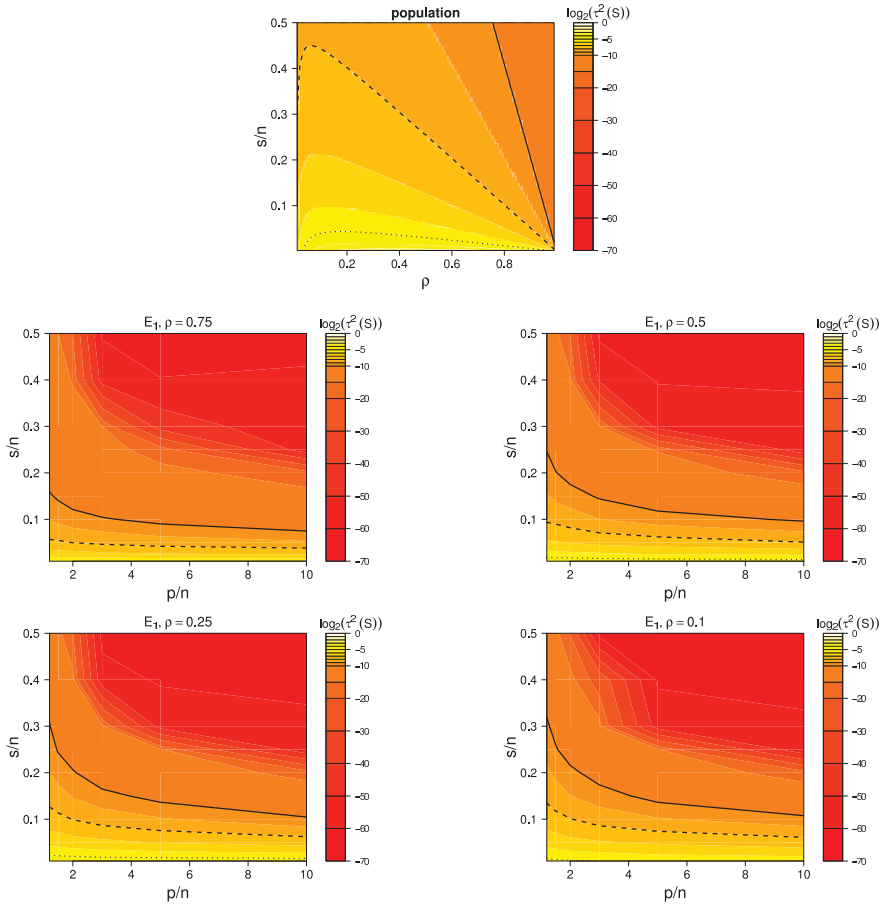

FIG 3. *Empirical scalings (0.05-quantiles over 100 replications) of the quantity* $\log_2(\tau^2(S))$ *for random design* E_1 *from the class* Ens_+ *in dependency of* s/n *and* p/n*, displayed in form of a contour plot. The lines indicates the level set for* -10 *(solid,* $2^{-10} \approx 0.001$*),* -8 *(dashed,* $2^{-8} \approx 0.004$*) and* -5 *(dotted,* $2^{-5} \approx 0.03$*). The top plot displays* $\log_2(\tau^2(S))$ *for the population Gram matrix in dependency of* s/n *and* $\rho \in (0, 1)$*.*

sets of 100 replications. It is revealed that for $\tau^2(S)$ to be positive, n does not need to be as large relative to s as suggested by Proposition 4. In fact, even for s/n as large as 0.3, $\tau^2(S)$ is sufficiently bounded away from zero as long as p is not dramatically larger than n ($p/n = 10$).

Implications for compressed sensing-type problems. In compressed sensing (CS) [13, 15], the goal is to recover a sparse vector β^* from a limited number of non-adaptive measurements. In the typical setup, the measurements are linear combinations of β^*, contaminated with additive noise and thus fall under the linear model (1.1). CS is related to *group testing* [20, 21]. Here, $\beta^* \in \{0, 1\}^p$ indicates the presence of a certain attribute of low prevalence in p objects, e.g. presence

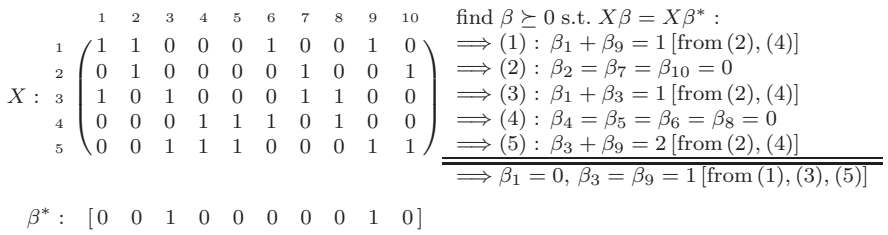

FIG 4. *Illustration of the setting. The rows of the sample measurement matrix X represent five group assignments yielding (noiseless) measurements $y = X\beta^*$. A NNLS estimator is thus a solution of the systems of linear equations $y = X\beta$ subject to $\beta \succeq 0$. A short calculation reveals that there is only a single solution β^* to this problem, even though it is severely underdetermined.*

of a rare disease in individuals or of a defect in manufactured goods. An effective strategy for locating the affected entities is by forming groups, testing for prevalence at the group level, discarding groups with a negative test result, and repeating the procedure with a new set of groups. The measurements obtained in this way are both adaptive and non-linear and hence do not fit into the conventional framework of CS. However, this is the case if it is possible to obtain aggregate measurements (i.e. sums) over arbitrary groups. As exemplified in Figure 4, the information of interest can be retrieved from few aggregated measurements and the associated group assignments. Proper measurement design needs to achieve a proper amount of overlapping of the groups. In fact, it is not possible to recover β^* from a reduced number of measurements involving only disjoint groups. At the same time, overlapping has to be limited to ensure that collections of $2s$ columns of the measurement matrix are linearly independent; otherwise, recovery of β^* is not possible in general. Without further prior knowledge, about the location of the non-zero entries of β^*, random group assignments in which each entity j, $j = 1, \ldots, p$, is assigned to group i, $i = 1, \ldots, n$, independently with probability π, appears to be reasonable in light of the results of the previous paragraph. Propositions 3 and 4 assert that, with high probability, the resulting measurement matrix suits well a sparse recovery approach based on NNLS.

Network tomography as discussed in [34] arises as a generalization of the above setting with measurements of the form $y = BA\beta^* + \varepsilon$, where $B \in \mathbb{R}_+^{n \times p}$ represents the measurement design and $A \in \mathbb{R}_+^{p \times p}$ is the adjacency matrix of the p nodes whose status of interest is contained in $\beta^* \in \mathbb{R}_+^p$. The goal is to spot sources of anomaly within the network. As distinguished from a random design setting, the measurement matrix $X = BA$ cannot be chosen freely, since A is fixed and the choice of B is subject to various constraints. Therefore, it is not clear a priori whether X satisfies our conditions, even though Σ is potentially self-regularizing as X is non-negative.

Sparsity of the solution. In Proposition 2, we have characterized the sparsity in the population setting. It is of interest to investigate this aspect for random de-

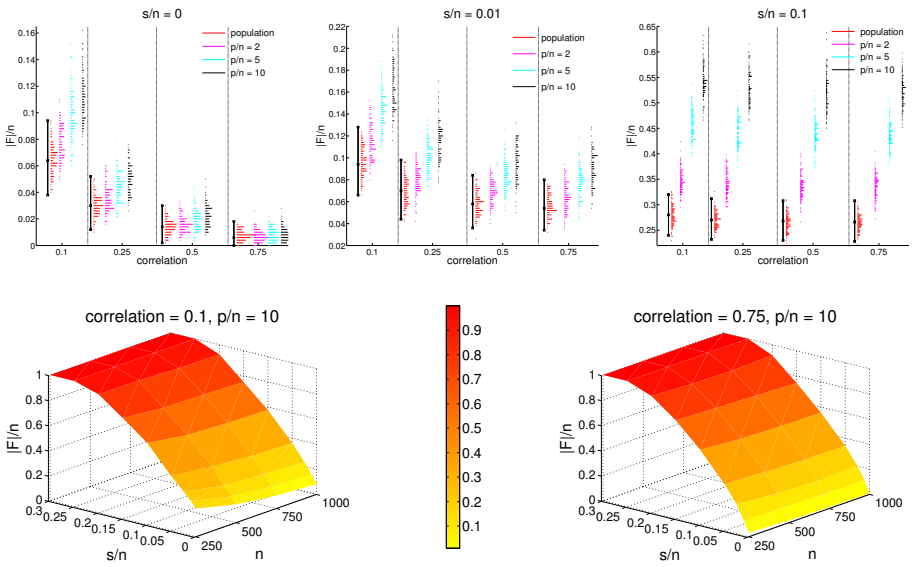

FIG 5. *Top: sparsity of the NNLS solution for random equi-correlation-like design in the $n < p$ setup as compared to the population setting. The squares represent the 0.01-, 0.5- and 0.99-quantiles of the (conditional) distribution of the fraction of active variables $|F|/n$ in the population according to Proposition 2. The vertical bars represent the empirical distributions over 100 random datasets with $n = 500$, where the colours correspond to different ratios p/n. Bottom: Surface plot of the 0.95-quantiles of $|F|/n$ over 100 random datasets for n and s/n varying.*

sign in the $p > n$ setup, particularly in light of Theorem 4, which implicitly relies on having a sparse NNLS solution. We here provide a sketch of the empirical behaviour within the experimental framework of the previous paragraph. We generate random design matrices ($n \in \{250, 500, 750, 1000\}$, $p/n \in \{2, 5, 10\}$) from E_1 for the four values of the parameter ρ as given above. For several values of s/n ranging from 0 to 0.3, we generate observations $y = X\beta^* + \varepsilon$, where ε is a Gaussian noise vector, and the components of β^*_S are set to the lower bound in Proposition 2. For each combination of $(n, p/n, s/n, \rho)$, 100 replications are considered and the fraction of active variables $|F|/n$ is determined.

Figure 5 summarizes the main findings of this experimental study. For fixed $n = 500$, the top panel depicts the empirical distributions of $|F|/n$ over the 100 replications in comparison to the population setting (cf. Figure 2). We observe that for all parameter configurations under consideration, the cardinalities of the active sets stay visibly away from 1 with $|F|/n$ being no larger than 2/3. The cardinalities of the active sets are larger than in the population case. The higher the sparsity level and the ratio p/n, the more pronounced the shifts toward larger cardinalities: while for $s/n = 0$ and $\rho = 0.75$, the empirical distribution of $|F|/n$ is rather close to that of the population, there is a consistent gap for $s/n = 0.1$. The bottom panel displays how $|F|/n$ scales with $(n, s/n)$. For plotting and space reasons, we restrict us to the 0.95-quantiles over the 100 replications and

$p/n = 10$, which, as indicated by the plots of the top panel, is the worst case among all values of p/n considered. The two surface plots for $\rho = 0.1$ and $\rho = 0.75$ are of a similar form; a noticeable difference occurs only for rather small s/n. It can be seen that for s/n fixed, $|F|/n$ roughly remains constant as n varies. On the other hand, $|F|/n$ increases rather sharply with s/n. For $s/n > 0.25$, we observe a breakdown, as $|F|/n = 1$. We point out that as long as $|F|/n < 1$, it holds that the NNLS solution and the active set are unique (with probability one), as follows from Lemma 5 in Appendix G.

6. Empirical performance

We here present the results of simulation studies in order to compare the performance of NNLS and the non-negative lasso in terms of prediction, estimation and sparse recovery.

6.1. Deconvolution of spike trains

We consider a positive spike-deconvolution model as in [30], as it commonly appears in various fields of applications. The underlying signal f is a function on $[a, b]$ of the form

$$f(u) = \sum_{k=1}^{s} \beta_k^* \phi_k(u),$$

with $\phi_k(\cdot) = \phi(\cdot - \mu_k)$, $k = 1, \ldots, s$, where $\phi \geq 0$ is given and the μ_k's define the locations of the spikes contained in $[a, b]$. The amplitudes $\{\beta_k^*\}_{j=1}^s$ are assumed to be positive. The goal is to determine the positions as well as the amplitudes of the spikes from n (potentially noisy) samples of the underlying signal f. As demonstrated below, NNLS can be a first step towards deconvolution. The idea is to construct a design matrix of the form $X = (\phi_j(u_i))$, where $\phi_j = \phi(\cdot - m_j)$ for candidate positions $\{m_j\}_{j=1}^p$ placed densely in $[a, b]$ and $\{u_i\}_{i=1}^n \subset [a, b]$ are the points at which the signal is sampled. Under an additive noise model with zero-mean sub-Gaussian noise ε, i.e.

$$y_i = \sum_{k=1}^{s} \beta_k^* \phi_k(u_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6.1}$$

and if X has the self-regularizing property (cf. Section 2), it follows immediately from Theorem 1 that the ℓ_2-prediction error of NNLS is bounded as

$$\frac{1}{n}\|f - X\widehat{\beta}\|_2^2 \leq \mathcal{E}^* + C\sqrt{\frac{\log p}{n}}, \qquad \text{where } \{f_i = f(u_i)\}_{i=1}^n, \tag{6.2}$$

where $\mathcal{E}^* = \min_{\beta \geq 0} \frac{1}{n}\|f - X\beta\|_2^2$. Even though it is not realistic to assume that $\{\mu_k\}_{k=1}^s \subset \{m_j\}_{j=1}^p$, i.e. that the linear model is correctly specified, we may think of $\mathcal{E}^*$ being negligible as long as the $\{m_j\}_{j=1}^p$ are placed densely enough.

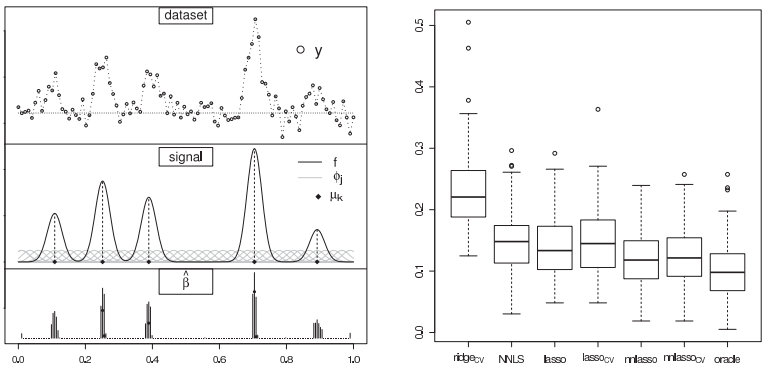

FIG 6. *Left panel: Visualization of the experimental setting. The middle part depicts the underlying signal, a positive combination of five Gaussians. The upper part depicts a sample dataset generated according to model* (6.1). *The lower part provides a summary of the coefficient vectors* $\widehat{\beta}$ *returned by NNLS, the heights of the bars representing the 0.9-quantiles and the dots the non-zero median coefficients at the respective positions over 100 replications. Right panel: Boxplots of the mean squared prediction errors (MSEs).*

This means that NNLS may be suitable for de-noising. Furthermore, the bound (6.2) implies that $\widehat{\beta}$ must have large components only for those columns of X corresponding to locations near the locations $\{\mu_k\}_{k=1}^s$ of the spikes, which can then be estimated accurately by applying a simple form of post-processing as discussed in [42]. On the other hand, the application of fast rate bounds such as that of Theorem 2 or corresponding results for the lasso is not adequate here, because the dense placement of the $\{\phi_j\}_{j=1}^p$ results into a tiny, if not zero, value of the restricted eigenvalue of Condition 2. For our simulation study, we consider model (6.1) as starting point. The signal is composed of five spikes of amplitudes between 0.2 and 0.7 convolved with a Gaussian function. The design matrix $X = (\phi_j(u_i))$ contains evaluations of $p = 200$ Gaussians $\{\phi_j\}_{j=1}^p$ at $n = 100$ points $\{u_i\}_{i=1}^n$, where both the centers $\{m_j\}_{j=1}^p$ of the $\{\phi_j\}_{j=1}^p$ as well as the $\{u_i\}_{i=1}^n$ are equi-spaced in the unit interval. We have by construction that $\{m_j\}_{j=1}^p \supset \{\mu_k\}_{k=1}^s$ so that $\mathcal{E}^* = 0$. The standard deviation of the Gaussians is chosen such that it is roughly twice the spacing of the $\{u_i\}$. At this point, it is important to note that the larger the standard deviations of the Gaussians, the larger the constant τ_0^2 (2.2), which here evaluates as $\tau_0^2 = 0.2876$. According to that setup, we generate 100 different vectors y resulting from different realizations of the noise ε whose entries are drawn i.i.d. from a Gaussian distribution with standard deviation $\sigma = 0.09$. The left panel of Figure 6 visualizes the setting. We compare the performance of NNLS, lasso/non-negative lasso with (i) fixed regularization parameter λ fixed to $\lambda_0 = 2\sigma\sqrt{2\log(p)/n}$ (ii) λ chosen from the grid $\lambda_0 \cdot 2^k, k = -5, -4, \ldots, 4$ by tenfold cross-validation, ridge regression (tuned by tenfold cross-validation) and an oracle least squares estimator based on knowledge of the positions $\{\mu_k\}_{k=1}^s$ of the spikes. The right panel of 6 contains

boxplots of the MSEs $\frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_2^2$ over all 100 replications. The performance of NNLS is only slightly inferior to that of the non-negative lasso, which is not far from the oracle, and roughly as good as that of the lasso. All methods improve substantially over ridge regression. The lower part of the left panel provides a summary of the NNLS estimator $\widehat{\beta}$, which is remarkably sparse and concentrated near the positions of the underlying spikes.

6.2. Sparse recovery

We now present the results of simulations in which we investigate the performance of NNLS with regard to estimation and sparse recovery in comparison to that of the non-negative lasso.

Setup. We generate data $y = X\beta^* + \varepsilon$, where ε has i.i.d. standard Gaussian entries. For the design X, two setups are considered.

Design I: Equi-correlation like design. The matrix X is generated by drawing its entries independently from the uniform distribution on $[0, 1]$ and rescaling them such that the population matrix is of equi-correlation structure with $\rho = 3/4$. Random matrices of that form have been considered for the numerical results in Section 5.3 (cf. Figures 2 to 5). For given (n, p, s), the target β^* is generated by setting $\beta_j^* = 6b \cdot \phi_{\min}^{-1/2}\sqrt{2\log(p)/n}(1 + U_j)$, $j = 1, \ldots, s$, where $\phi_{\min} = (1 - \rho)$ denotes the smallest eigenvalue of the population Gram matrix, the $\{U\}_{j=1}^s$ are drawn uniformly from $[0, 1]$, and we let the parameter $b > 0$ vary. We further set $\beta_j^* = 0$, $j = (s + 1), \ldots, p$.

Design II: Localized non-negative functions. The setup leading to the second class of designs can be regarded as a simplification of the deconvolution problem discussed in the previous subsection to fit into the standard sparse recovery framework. Indeed, in the experiments of Section 6.1, recovery of the support of β^* fails in the presence of noise, because the $\{\phi_j\}$'s are placed too densely relative to the number of sampling points; see [11] for a similar discussion concerning the recovery of mixture components in sparse mixture density estimation. In order to circumvent this issue, we use the following scheme. As in Section 6.1, we consider sampling points $u_i = i/n$, $i = 1, \ldots, n$, in $[0, 1]$ and localized functions $\phi_j = \phi(\cdot - m_j)$,where here $\phi(\cdot - m_j) = \exp(-|\cdot - m_j|/h)$, $j = 1, \ldots, p$ with $h = 2/n$. The centers m_j, $j = 1, \ldots, p$, are taken from the interval $[m_{\min}, m_{\max}]$, where $m_{\min} = u_1 - h\log(1/n)$ and $m_{\max} = u_n + h\log(1/n)$. Given the sparsity level s, $[m_{\min}, m_{\max}]$ is partitioned into s sub-intervals of equal length and the centers $m_1, \ldots, m_s$ corresponding to S are drawn from the uniform distributions on these intervals. The remaining centers $m_{s+1}, \ldots, m_p$ corresponding to S^c are drawn uniformly from $[m_{\min}, m_{\max}] \setminus \cup_{j=1}^s [m_j - \Delta, m_j + \Delta]$, where $\Delta > 0$ is set to enforce a sufficient amount of separation of the $\{\phi_j\}_{j=1}^s$ from the $\{\phi_j\}_{j=s+1}^p$. We here choose $\Delta = h = 2/n$. The design matrix is then of the form $X_{ij} = \phi_j(u_i)/c_j$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, where the c_j's are scaling factors such that $\|X_j\|_2^2 = n\,\forall j$. As for Design I, we generate observations $y = X\beta^* + \varepsilon$, where

$\beta_j^* = b \cdot \beta_{\min}(1 + U_j)$, $j = 1, \ldots, s$ and $\beta_j^* = 0$, $j = s+1, \ldots, p$. The $\{U_j\}_{j=1}^s$ are random variables from the uniform distribution on $[0, 1]$ and the choice $\beta_{\min} = 4\sqrt{6\log(10)/n}$ has turned out to yield sufficiently challenging problems.

For both Design I and II, two sets of experiments are performed. In the first one, the parameter b controlling the magnitude of the coefficients of the support is fixed to $b = 0.5$ (Design I) respectively $b = 0.55$ (Design II), while the aspect ratio p/n of X and the fraction of sparsity s/n vary. In the second set of experiments, s/n is fixed to 0.2 (Design I) and 0.05 (Design II), while p/n and b vary. Each configuration is replicated 100 times for $n = 500$.

Comparison. Across these runs, we compare thresholded NNLS, the non-negative lasso (NNℓ_1), the thresholded non-negative lasso (tNNℓ_1) and orthogonal matching pursuit (OMP, [49, 59]) with regard to their performance in sparse recovery. Additionally, we compare NNLS and NNℓ_1 with $\lambda = \lambda_0$ as defined below (both *without* subsequent thresholding) with regard to estimation of β^* in ℓ_∞-norm (Tables 1 and 2) and ℓ_2-norm (the results are contained in the supplement [45]). The performance of thresholded NNLS with regard to sparse recovery is assessed in two ways. For the first one (referred to as 'tNNLS*'), success is reported whenever $\min_{j\in S} \widehat{\beta}_j > \max_{j\in S^c} \widehat{\beta}_j$, i.e. there exists a threshold that permits support recovery. Second, the procedure of Theorem 5 (with σ replaced by the naive estimate $\frac{1}{n}\|y - X\widehat{\beta}\|_2^2$) is used to determine the threshold in a data-driven manner without knowledge of S. This approach is referred to as tNNLS. For tNNℓ_1, both the regularization parameter λ and the threshold have to be specified. Instead of fixing λ to a single value, we give tNNℓ_1 a slight advantage by simultaneously considering all solutions $\lambda \in [\lambda_0 \wedge \widehat{\lambda}, \lambda_0 \vee \widehat{\lambda}]$ prior to thresholding, where $\lambda_0 = 2\sigma\sqrt{2\log(p)/n}$ equals the choice of the regularization parameter advocated in [5] to achieve the optimal rate for the estimation of β^* in the ℓ_2-norm and $\widehat{\lambda} = 2\|X^\top \varepsilon/n\|_\infty$ can be interpreted as empirical counterpart to λ_0. The non-negative lasso modification of LARS [22] is used to obtain the solutions $\{\widehat{\beta}(\lambda) : \lambda \in [\lambda_0 \wedge \widehat{\lambda}, \lambda_0 \vee \widehat{\lambda}]\}$; we then report success of tNNℓ_1 whenever $\min_{j\in S} \widehat{\beta}_j(\lambda) > \max_{j\in S^c} \widehat{\beta}_j(\lambda)$ holds for at least *one* of these solutions. We point out that specification of λ_0 is based on knowledge of the noise variance, which constitutes a second potential advantage for tNNℓ_1.

Under the conditions of Theorem 6, NNℓ_1 recovers the support directly without thresholding. In order to judge the usefulness of subsequent thresholding of NNℓ_1, we obtain as well the set of non-negative lasso solutions $\{\widehat{\beta}(\lambda) : \lambda \geq \lambda_0 \wedge \widehat{\lambda}\}$ and check whether the sparsity pattern of any of these solutions recovers S.

Given its simplicity, OMP serves as basic reference method. Success is reported whenever the support has been recovered after s steps.

Discussion of the results. In summary, Figures 7 and 8 indicate that for the two setups under consideration, NNLS and its thresholded version exhibit excellent performance in sparse recovery. A superior performance relative to the thresholded non-negative lasso is achieved particularly in more difficult parameter regimes characterized by comparatively small signal strength b or high

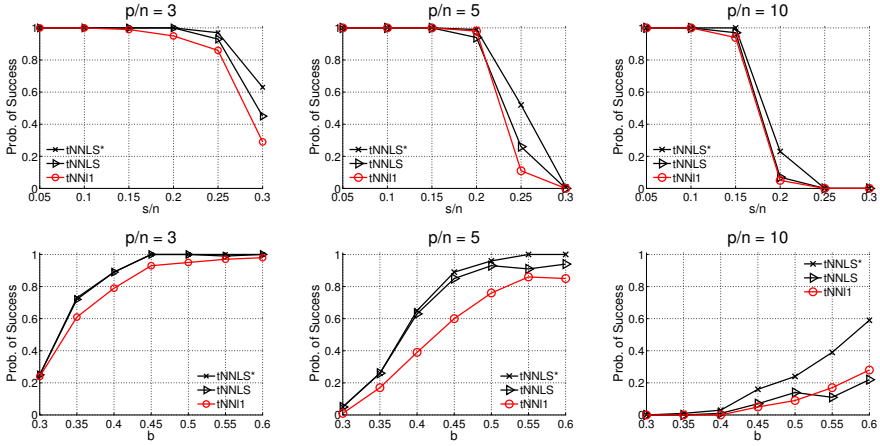

FIG 7. *Sparse recovery results for Design I. Top: Results of the set of experiments with fixed signal strength* $b = 0.5$. *Bottom: Results of the set of experiments with fixed fraction of sparsity* $s/n = 0.2$. *'tNNLS*' and 'tNNℓ_1' denote thresholded NNLS and the thresholded non-negative lasso, where thresholding is done with knowledge of* S. *'tNNLS' denotes thresholded NNLS with data-driven choice of the threshold. The results of the non-negative lasso without thresholding and OMP are not displayed, because these two approaches fail in all instances.*

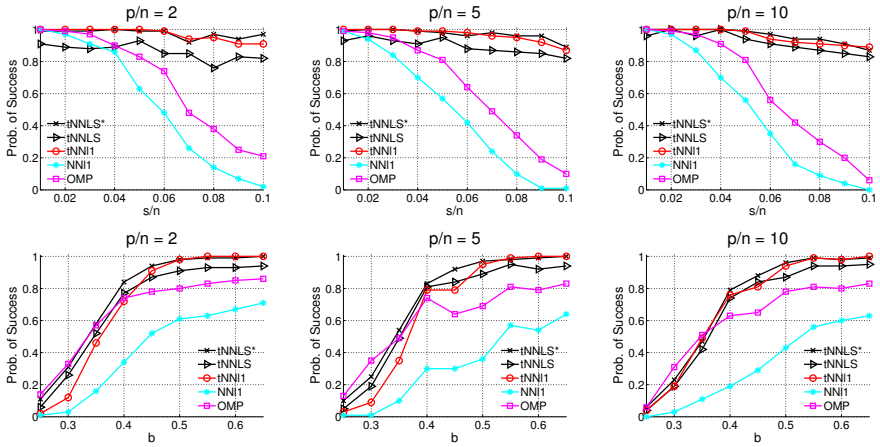

FIG 8. *Sparse recovery results for Design II. Top: Results of the set of experiments with fixed signal strength* $b = 0.55$. *Bottom: Results of the set of experiments with fixed fraction of sparsity* $s/n = 0.05$. *'tNNLS*' and 'tNNℓ_1' denote thresholded NNLS and the thresholded non-negative lasso, where thresholding is done with knowledge of* S. *'tNNLS' denotes thresholded NNLS with data-driven choice of the threshold. 'NNℓ_1' denotes the non-negative lasso without thresholding and 'OMP' orthogonal matching pursuit.*

fraction of sparsity. The results of the experiments reveal that the non-negative lasso without thresholding may perform well in estimation, but it is not competitive as far as sparse recovery is concerned. This observation is in agreement

TABLE 1
Average sup-norm errors (± standard errors) of $\|\widehat{\beta} - \beta^*\|_\infty$ (NNLS) and $\|\widehat{\beta}^{\ell_1,\geq} - \beta^*\|_\infty$ (NNℓ_1) for Design I with $b = 0.5$

	p/n			
	2		3	
s/n	nnls	nnℓ_1	nnls	nnℓ_1
0.05	.34±.005	.34±.005	.35±.005	.36±.005
0.1	.37±.005	.37±.005	.41±.005	.40±.005
0.15	.41±.006	.42±.009	.44±.005	.46±.012
0.2	.43±.006	.46±.012	.50±.007	.56±.023
0.25	.48±.006	.54±.020	.58±.009	.72±.030
0.3	.55±.007	.64±.027	.70±.012	1.01±.04
	p/n			
	5		10	
s/n	nnls	nnℓ_1	nnls	nnℓ_1
0.05	.37±.005	.38±.005	.43±.006	.43±.006
0.1	.44±.005	.45±.006	.51±.007	.52±.007
0.15	.52±.007	.54±.007	.66±.009	.71±.012
0.2	.61±.008	.66±.009	1.01±.02	1.28±.03
0.25	.81±.014	1.32±.04	1.91±.02	2.17±.02
0.3	1.36±.03	1.90±.03	2.32±.02	2.36±.03

TABLE 2
Average sup-norm errors (± standard errors) of $\|\widehat{\beta} - \beta^*\|_\infty$ (NNLS) and $\|\widehat{\beta}^{\ell_1,\geq} - \beta^*\|_\infty$ (NNℓ_1) for Design II with $b = 0.55$

	p/n			
	2		3	
s/n	nnls	nnℓ_1	nnls	nnℓ_1
0.02	.20±.005	.32±.005	.21±.005	.32±.005
0.04	.23±.004	.34±.004	.24±.007	.35±.006
0.06	.25±.006	.36±.005	.27±.008	.37±.006
0.08	.28±.010	.37±.009	.28±.009	.37±.006
0.1	.29±.010	.37±.007	.32±.012	.39±.010
	p/n			
	5		10	
s/n	nnls	nnℓ_1	nnls	nnℓ_1
0.02	.21±.004	.32±.004	.22±.005	.33±.006
0.04	.23±.005	.34±.004	.24±.005	.35±.005
0.06	.27±.005	.36±.005	.27±.006	.37±.006
0.08	.29±.011	.37±.009	.30±.009	.37±.006
0.1	.32±.011	.40±.010	.32±.011	.39±.008

with existing literature in which the restrictiveness of the conditions for the lasso to select the correct set of variables is pointed out and two stage procedures like thresholding are proposed as remedy [36, 50, 60, 63, 64]. At this point, we stress again that NNLS only requires one parameter (the threshold) to be set,

whereas competitive performance with regard to sparse recovery based on the non-negative lasso entails specification of two parameters. Let us now give some more specific comments separately for the two designs. For Design I, thresholded NNLS visibly improves over tNNℓ_1, predominantly even in case that the threshold is chosen adaptively without knowledge of S. For Design II, noticeable differences between tNNLS* and tNNℓ_1 occur for small values of b. Apart from that, the performance is essentially identical. Even though the results of tNNLS remain competitive, they fall behind those tNNLS* and tNNℓ_1. OMP partially keeps up with the other methods for s/n and/or b small, while NNℓ_1 succeeds as well in a substantial fraction of cases for small s/n. This is to be contrasted with the situation for Design I, in which both OMP and NNℓ_1 do not even achieve success in a single trial. This outcome is a consequence of the fact that the non-negative irrepresentable condition (cf. Section 4.4), which is necessary for the success of OMP as well [59], fails to hold in all these runs. The ℓ_∞-errors in estimating β^* reported in Tables 1 and 2 are in accordance with the sparse recovery results. The smaller s/n and p/n, the closer NNLS and NNℓ_1 are in performance. An advantage of NNLS arises for more extreme combinations of $(s/n, p/n)$. A similar conclusion can be drawn for the ℓ_2-errors (cf. supplement [45]).

Acknowledgements

We would like to thank Mike Davies, Rolf Schneider, Jared Tanner and Roman Vershynin for helpful discussions. We thank the Editor and the reviewers for thorougly reading the manuscript and providing valuable comments and suggestions. We are also indebted to a reviewer of an earlier draft, whose comments prompted us to organize Section 3 in its present form.

Appendix A: Standard tail bounds for sub-Gaussian random variables

A zero-mean random variable Z is called sub-Gaussian if there exists $\sigma > 0$ (referred to as sub-Gaussian parameter) so that the moment-generating function obeys the bound $\mathbf{E}[\exp(tZ)] \leq \exp(\sigma^2 t^2/2),\ \forall t \in \mathbb{R}$. It follows that if $Z_1, \ldots, Z_n$ are i.i.d. copies of Z and $v \in \mathbb{R}^n$, then $\sum_{i=1}^n v_i Z_i$ is sub-Gaussian with parameter $\|v\|_2^2 \sigma^2$. We have the well-known tail bound

$$\mathbf{P}(|Z| > z) \leq 2 \exp\left(-\frac{z^2}{2\sigma^2}\right), \quad z \geq 0. \tag{A.1}$$

Combining the previous two facts and using a union bound with $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$ it follows that for any collection of vectors $v_j \in \mathbb{R}^n$, $j = 1, \ldots, p$,

$$\mathbf{P}\left(\max_{1 \leq j \leq p} |v_j^\top \mathbf{Z}| > \sigma \max_{1 \leq j \leq p} \|v_j\|_2 \left(\sqrt{2 \log p} + z\right)\right) \leq 2 \exp\left(-\frac{1}{2} z^2\right), \quad z \geq 0. \tag{A.2}$$

To obtain the main results of the paper, (A.2) is applied with $Z = \varepsilon$ and $v_j = X_j/n$ under the assumption $\|X_j\|_2^2 = n$, $j = 1, \ldots, p$, $j = 1, \ldots, p$, and the choice $z = M\sqrt{2\log p}$ for $M \geq 0$, which yields

$$\mathbf{P}\left(\max_{1\leq j\leq p}\left|\frac{X_j^\top \varepsilon}{n}\right| > \sigma(1+M)\sqrt{\frac{2\log p}{n}}\right) \leq 2p^{-M^2}. \tag{A.3}$$

Appendix B: Proof of Proposition 1

Since X satisfies Condition 1, by (2.2), there exists a unit vector w so that

$$\frac{X^\top w}{\sqrt{n}} = h, \quad \text{where } h \succeq \tau \mathbf{1}, \tag{B.1}$$

for some constant $\tau > 0$. Setting $\Pi = I - ww^\top$ as the projection on the subspace orthogonal to w, the least squares objective can be decomposed as follows.

$$\begin{aligned}
\frac{1}{n}\|\varepsilon - X\beta\|_2^2 &= \frac{\varepsilon^\top\varepsilon}{n} - \frac{2\varepsilon^\top X\beta}{n} + \frac{\beta^\top X^\top X\beta}{n} \\
&= \left(\frac{\varepsilon^\top\varepsilon}{n} - \frac{2\varepsilon^\top \Pi X\beta}{n} + \frac{\beta^\top X^\top \Pi X\beta}{n}\right) + \frac{\beta^\top X^\top ww^\top X\beta}{n} \\
&\quad - \frac{2\varepsilon^\top ww^\top X\beta}{n} \\
&= \frac{1}{n}\|\varepsilon - \Pi X\beta\|_2^2 + (h^\top\beta)^2 - \frac{2\varepsilon^\top w}{\sqrt{n}}h^\top\beta \\
&= \frac{1}{n}\|\varepsilon - \overline{X}\beta\|_2^2 + (h^\top\beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right)h^\top\beta
\end{aligned}$$

where $\overline{X} = \Pi X$. In the last line, we have invoked the assumptions made for ε. Writing H for the diagonal matrix with the entries of h/τ on its diagonal and setting $D = H^{-1}$ and $\widetilde{X} = \overline{X}D = (X\Pi)D$, we have

$$\begin{aligned}
&\min_{\beta\succeq 0}\frac{1}{n}\|\varepsilon - \overline{X}\beta\|_2^2 + (h^\top\beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right)h^\top\beta \\
=&\min_{\beta\succeq 0}\frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2(\mathbf{1}^\top\beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right)\tau\mathbf{1}^\top\beta,
\end{aligned}$$

where we have used (B.1). Note that by (B.1) and $\tau \leq 1$, D has the property claimed in the statement. In view of the presence of the term $\tau^2(\mathbf{1}^\top\beta)^2$, any minimizer β° of the r.h.s. must obey $\mathbf{1}^\top\beta^\circ = O_{\mathbf{P}}(1)$. As a result,

$$\begin{aligned}
&\min_{\beta\succeq 0}\frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2(\mathbf{1}^\top\beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right)\tau\mathbf{1}^\top\beta \\
=&\min_{\beta\succeq 0}\frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2(\mathbf{1}^\top\beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

which finishes the proof of the first claim of the proposition. To establish the second claim, observe that any $\widehat{\beta} \in \operatorname{argmin}_{\beta \geq 0} \frac{1}{n}\|y - X\beta\|_2^2$ satisfies

$$\frac{1}{n}\|\varepsilon - X\widehat{\beta}\|_2^2 \leq \frac{1}{n}\|\varepsilon\|_2^2.$$

Expanding the square and re-arranging, we obtain

$$\frac{1}{n}\|X\widehat{\beta}\|_2^2 \leq \frac{2\varepsilon^\top X\widehat{\beta}}{n} \leq 2\frac{\|X^\top\varepsilon\|_\infty}{n}\mathbf{1}^\top\widehat{\beta}.$$

As established above, $\mathbf{1}^\top\widehat{\beta} = O_{\mathbf{P}}(1)$, so that $\frac{1}{n}\|X\widehat{\beta}\|_2^2 = o_{\mathbf{P}}(1)$ as long as $\frac{1}{n}\|X^\top\varepsilon\|_\infty = o_{\mathbf{P}}(1)$.

Appendix C: Proof of Theorem 1

Since $\widehat{\beta}$ is a minimizer of the NNLS problem (1.2) and since β^* is a feasible solution, we have that

$$\begin{aligned}
&\frac{1}{n}\|y - X\widehat{\beta}\|_2^2 \leq \frac{1}{n}\|y - X\beta^*\|_2^2 \\
\Leftrightarrow\ &\frac{1}{n}\|(f + \varepsilon - X\beta^*) + X\beta^* - X\widehat{\beta}\|_2^2 \leq \frac{1}{n}\|f + \varepsilon - X\beta^*\|_2^2 \\
\Rightarrow\ &\frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_2^2 + \frac{2}{n}(f + \varepsilon - X\beta^*)^\top X(\beta^* - \widehat{\beta}) \leq 0 \\
\Rightarrow\ &\frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_2^2 \leq \frac{2}{n}(f - X\beta^*)^\top X(\widehat{\beta} - \beta^*) + \frac{2}{n}\varepsilon^\top X(\widehat{\beta} - \beta^*).
\end{aligned} \tag{C.1}$$

Write $\widehat{\delta} = \widehat{\beta} - \beta^*$, $P = \{j : \widehat{\delta}_j \geq 0\}$ and $N = \{j : \widehat{\delta}_j < 0\}$. We now lower bound $\frac{1}{n}\|X\widehat{\delta}\|_2^2 = \widehat{\delta}^\top\Sigma\widehat{\delta}$ using the self-regularizing property of Condition 1 according to (2.3).

$$\begin{aligned}
\frac{1}{n}\|X\widehat{\delta}\|_2^2 &= \widehat{\delta}_P^\top\Sigma_{PP}\widehat{\delta}_P + 2\widehat{\delta}_P^\top\Sigma_{PN}\widehat{\delta}_N + \widehat{\delta}_N^\top\Sigma_{NN}\widehat{\delta}_N \\
&\geq \tau^2(\mathbf{1}^\top\widehat{\delta}_P)^2 - 2\|\widehat{\delta}_P\|_1\|\widehat{\delta}_N\|_1.
\end{aligned} \tag{C.2}$$

Second, we bound the r.h.s. of (C.1). We set $A = \max_{1\leq j\leq p}\left|\frac{1}{n}X_j^\top\varepsilon\right|$ and use the bound

$$\max_{1\leq j\leq p}\left|\frac{1}{n}X_j^\top(f - X\beta^*)\right| \leq \max_{1\leq j\leq p}\frac{1}{\sqrt{n}}\|X_j\|_2\sqrt{\frac{1}{n}\|f - X\beta^*\|_2^2} = \sqrt{\mathcal{E}^*},$$

obtaining that

$$\frac{1}{n}\|X\widehat{\delta}\|_2^2 \leq 2(A + \sqrt{\mathcal{E}^*})\|\widehat{\delta}\|_1 \tag{C.3}$$

Inserting the lower bound (C.2) into (C.3), we obtain

$$\tau^2\|\widehat{\delta}_P\|_1^2 - 2\|\widehat{\delta}_P\|_1\|\widehat{\delta}_N\|_1 \leq 2(A + \sqrt{\mathcal{E}^*})(\|\widehat{\delta}_P\|_1 + \|\widehat{\delta}_N\|_1). \tag{C.4}$$

We may assume that $\widehat{\delta}_P \neq 0$, otherwise the assertion of the theorem would follow immediately, because $\|\widehat{\delta}_N\|_1$ is already bounded for feasibility reasons, see below. Dividing both sides by $\|\widehat{\delta}_P\|_1$ and re-arranging yields

$$\|\widehat{\delta}_P\|_1 \leq \frac{4(A + \sqrt{\mathcal{E}^*}) + 2\|\widehat{\delta}_N\|_1}{\tau^2}, \tag{C.5}$$

where we have assumed that $\|\widehat{\delta}_N\|_1 \leq \|\widehat{\delta}_P\|_1$ (if that were not the case, one would obtain $\|\widehat{\delta}_P\|_1 \leq \|\widehat{\delta}_N\|_1$, which is stronger than (C.5), since $0 < \tau^2 \leq 1$). We now substitute (C.5) back into (C.1) and add $\mathcal{E}^* = \frac{1}{n}\|X\beta^* - f\|_2^2$ to both sides of the inequality in order to obtain

$$\begin{aligned}
\widehat{\mathcal{E}} = \frac{1}{n}\|X\widehat{\beta} - f\|_2^2 &\leq \mathcal{E}^* + 2A(\|\widehat{\delta}_P\|_1 + \|\widehat{\delta}_N\|_1) \\
&\leq \mathcal{E}^* + 2A\left(\frac{4(A + \sqrt{\mathcal{E}^*}) + 2\|\widehat{\delta}_N\|_1}{\tau^2} + \|\widehat{\delta}_N\|_1\right) \\
&\leq \mathcal{E}^* + \frac{6A\|\beta^*\|_1 + 8(A^2 + A\sqrt{\mathcal{E}^*})}{\tau^2},
\end{aligned}$$

noting that by feasibility of $\widehat{\beta}$, one has $\widehat{\delta} \succeq -\beta^*$ and hence $\|\widehat{\delta}_N\|_1 \leq \|\beta^*\|_1$. Using (A.3), the event $\{A \leq (1+M)\sigma\sqrt{\frac{2\log p}{n}}\}$ holds with probability no less than $1 - 2p^{-M^2}$. The result follows.

Appendix D: Proofs of Lemma 1 and Theorem 2

We build on ideas already used in the proof of Theorem 1. In particular, all notations introduced in the previous proof are adopted. First note that $S^c \subseteq P$ and $N \subseteq S$. Hence, we obtain the following analog to (C.4).

$$\tau^2\|\widehat{\delta}_{S^c}\|_1^2 - 2\|\widehat{\delta}_{S^c}\|_1\|\widehat{\delta}_S\|_1 \leq 2A(\|\widehat{\delta}_S\|_1 + \|\widehat{\delta}_{S^c}\|_1).$$

Dividing both sides by $\|\widehat{\delta}_{S^c}\|_1$, assuming that $0 < \|\widehat{\delta}_S\|_1 \leq \|\widehat{\delta}_{S^c}\|_1$ (otherwise, the claim $\widehat{\delta} \in \mathcal{R}(3/\tau^2, S)$ as in the first event of Lemma 1 would follow trivially), we obtain

$$\tau^2\|\widehat{\delta}_{S^c}\|_1 \leq 4A + 2\|\widehat{\delta}_S\|_1.$$

If $4A \leq \|\widehat{\delta}_S\|_1$, then the first event of Lemma 1 occurs. Otherwise, we conclude that the second event of Lemma 1 occurs by applying (A.3) to bound A as in the proof of Theorem 1.

In the latter case, the assertion of Theorem 2 follows immediately. Thus, the rest of the proof is conditional on the first event. In terms of Condition 2, $\widehat{\delta} \in \mathcal{R}(3/\tau^2, S)$, so that one may invoke the restricted eigenvalue condition (3.2), which, when applied to (C.1), yields

$$\phi\left(\frac{3}{\tau^2}, s\right)\|\widehat{\delta}_S\|_2^2 \leq \frac{1}{n}\|X\widehat{\delta}\|_2^2 \leq 2A(\|\widehat{\delta}_S\|_1 + \|\widehat{\delta}_{S^c}\|_1) \leq 2\left(1 + \frac{3}{\tau^2}\right)A\|\widehat{\delta}_S\|_1$$

which implies that

$$\|\widehat{\delta}_S\|_1 \leq \frac{2s}{\phi\left(\frac{3}{\tau^2}, s\right)}\left(1+\frac{3}{\tau^2}\right)A \implies \|\widehat{\delta}\|_1 \leq \frac{2s}{\phi\left(\frac{3}{\tau^2}, s\right)}\left(1+\frac{3}{\tau^2}\right)^2 A$$

The preceding bound in turn implies

$$\frac{1}{n}\|X\widehat{\delta}\|_2^2 \leq \frac{4s}{\phi\left(\frac{3}{\tau^2}, s\right)}\left(1+\frac{3}{\tau^2}\right)^2 A^2.$$

Controlling A as above, the ℓ_1-bound and the bound on the prediction error follow. The bound on $\|\widehat{\delta}\|_q^q$ for $q \in (1,2]$ can be established similarly. The proof is along the lines of the proof of Theorem 7.2 in [5] and is hence omitted.

Appendix E: Proof of Lemma 2

The proof of the lemma relies on the following auxiliary result, which is immediate from the KKT optimality conditions of the NNLS problem. Its proof is hence omitted.

Lemma 3. *$\widehat{\beta}$ is a minimizer of the NNLS problem (1.2) if and only if there exists $F \subseteq \{1,\ldots,p\}$ such that*

$$\frac{1}{n}X_j^\top(y - X\widehat{\beta}) = 0, \text{ and } \widehat{\beta}_j > 0,\ j \in F, \quad \frac{1}{n}X_j^\top(y - X\widehat{\beta}) \leq 0, \text{ and } \widehat{\beta}_j = 0,\ j \in F^c.$$

Lemma 3 implies that any NNLS solution is a minimizer of a least squares problem subject to the equality constraint $\beta_{F^c} = 0$ given the active set F, that is

$$\frac{1}{n}\|y - X\widehat{\beta}\|_2^2 = \min_{\beta} \frac{1}{n}\|y - X\beta\|_2^2 \quad \text{subject to } \beta_{F^c} = 0.$$

Proof of Lemma 2.. The NNLS objective can be split into two parts as follows.

$$\frac{1}{n}\|y - X\beta\|_2^2 = \frac{1}{n}\|\Pi_S y - X_S\beta_S - \Pi_S X_{S^c}\beta_{S^c}\|_2^2 + \frac{1}{n}\|\xi - Z\beta_{S^c}\|_2^2, \qquad \text{(E.1)}$$

Separate minimization of the second summand on the r.h.s. of (E.1) yields $\widehat{\beta}^{(P1)}$. Substituting $\widehat{\beta}^{(P1)}$ for β_{S^c} in the first summand, and minimizing the latter amounts to solving $(P2)$. In view of Lemma 3, if $\widehat{\beta}^{(P2)} \succ 0$, it coincides with an unconstrained least squares estimator corresponding to problem $(P2)$. This implies that the optimal value of $(P2)$ must be zero, because the observation vector $X_S\beta_S^* + \Pi_S(\varepsilon - X_{S^c}\widehat{\beta}^{(P1)})$ of the non-negative least squares problem $(P2)$ is contained in the column space of X_S. Since the second summand in (E.1) corresponding to $(P1)$ cannot be made smaller than by separate minimization, we have minimized the non-negative least squares objective. □

Appendix F: Proof of Theorem 3

We here state and prove a result that is slightly more general than Theorem 3, as it covers also the case of an approximately sparse target. Writing $\beta_{(1)}^* \geq \cdots \geq$

$\beta^*_{(p)} \geq 0$ for the sequence of ordered coefficients, let $S = \{j : \beta^*_j \geq \beta^*_{(s)}\}$ be the set of the s largest coefficients of β^* (for simplicity, assume that there are no ties). For the result that follows, we think of $\|\beta^*_{S^c}\|_1$ being considerably smaller than the entries of β^*_S.

Theorem 7. *Consider the linear model $y = X\beta^* + \varepsilon$, where $\beta^* \succeq 0$ and ε has i.i.d. zero-mean sub-Gaussian entries with sub-Gaussian parameter σ. For $M \geq 0$, set*

$$b = \frac{2\left(\|\beta^*_{S^c}\|_1 + (1+M)\sigma\sqrt{\frac{2\log p}{n}}\right)}{\tau^2(S)},$$

$$\text{and } \widetilde{b} = (b + \|\beta^*_{S^c}\|_1) \cdot K(S) + \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(S)}}\sqrt{\frac{2\log p}{n}}.$$

If $\beta_{\min}(S) > \widetilde{b}$, then the NNLS estimator $\widehat{\beta}$ has the following properties with probability no less than $1 - 4p^{-M^2}$:

$$\|\widehat{\beta}_{S^c}\|_1 \leq b \quad \text{and} \quad \|\widehat{\beta}_S - \beta^*_S\|_\infty \leq \widetilde{b}.$$

The special case of exact sparsity in which S equals the support of β^* is obtained for $\|\beta^*_{S^c}\|_1 = 0$ (cf. Theorem 3).

Proof. First note that in the more general case with $\beta^*_{S^c} \neq 0$, an analog of Lemma 2 holds with

$$(P1): \min_{\beta^{(P1)} \succeq 0} \frac{1}{n}\|\xi + Z\beta^*_{S^c} - Z\beta^{(P1)}\|_2^2,$$

$$(P2): \min_{\beta^{(P2)} \succeq 0} \frac{1}{n}\|\Pi_S\varepsilon + \Pi_S X_{S^c}\beta^*_{S^c} + X_S\beta^*_S - X_S\beta^{(P2)} - \Pi_S X_{S^c}\widehat{\beta}^{(P1)}\|_2^2$$

Consider problem $(P1)$.

Step 1: *Controlling $\|\widehat{\beta}^{(P1)}\|_1$ via $\tau^2(S)$.* Since $\widehat{\beta}^{(P1)}$ is a minimizer and 0 is feasible for $(P1)$, we have

$$\frac{1}{n}\|\xi + Z\beta^*_{S^c} - Z\widehat{\beta}^{(P1)}\|_2^2 \leq \frac{1}{n}\|\xi + Z\beta^*_{S^c}\|_2^2,$$

which implies that

$$\begin{aligned}
(\widehat{\beta}^{(P1)})^\top \frac{1}{n} Z^\top Z\widehat{\beta}^{(P1)} &\leq \|\widehat{\beta}^{(P1)}\|_1 \left(A + 2\left\|\tfrac{1}{n}Z^\top Z\beta^*_{S^c}\right\|_\infty\right), \quad A = \max_j \frac{2}{n}|Z_j^\top \xi|. \\
&\leq \|\widehat{\beta}^{(P1)}\|_1 \left(A + 2\|\beta^*_{S^c}\|_1 \max_{j,k} Z_j^\top Z_k/n\right) \\
&\leq \|\widehat{\beta}^{(P1)}\|_1 \left(A + 2\max_{j,k}\|Z_j/\sqrt{n}\|_2\|Z_k/\sqrt{n}\|_2\|\beta^*_{S^c}\|_1\right) \\
&\leq \|\widehat{\beta}^{(P1)}\|_1 (A + 2\|\beta^*_{S^c}\|_1).
\end{aligned} \tag{F.1}$$

In the last inequality, we have used that for all $j = 1, \ldots, p$, it holds that

$$\|Z_j\|_2 = \|\Pi_S^{\perp} X_j\|_2 \leq \|X_j\|_2. \tag{F.2}$$

As observed in (4.3), $\tau^2(S) = \min_{\lambda \in T^{p-s-1}} \lambda^\top \frac{1}{n} Z^\top Z \lambda$, s.t. the l.h.s. of (F.1) can be lower bounded via

$$(\widehat{\beta}^{(P1)})^\top \frac{1}{n} Z^\top Z \widehat{\beta}^{(P1)} \geq \tau^2(S) \|\widehat{\beta}^{(P1)}\|_1^2. \tag{F.3}$$

Combining (F.1) and (F.3), we have $\|\widehat{\beta}^{(P1)}\|_1 \leq (A + 2\|\beta^*_{S^c}\|_1)/\tau^2(S)$.

Step 2: *Back-substitution into (P2).* Equipped with the bound just derived, we insert $\widehat{\beta}^{(P1)}$ into problem $(P2)$ of Lemma 2, and show that in conjunction with the assumptions made for the minimum support coefficient $\beta_{\min}(S)$, the *ordinary* least squares estimator corresponding to $(P2)$

$$\bar{\beta}^{(P2)} = \operatorname*{argmin}_{\beta^{(P2)}} \frac{1}{n} \|\Pi_S y - X_S \beta^{(P2)} - \Pi_S X_{S^c} \widehat{\beta}^{(P1)}\|_2^2$$

has only positive components. Lemma 2 then yields $\bar{\beta}^{(P2)} = \widehat{\beta}^{(P2)} = \widehat{\beta}_S$. Using the closed form expression for the ordinary least squares estimator, one obtains

$$\begin{aligned} \bar{\beta}^{(P2)} &= \frac{1}{n} \Sigma_{SS}^{-1} X_S^\top \Pi_S (y - X_{S^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})) \\ &= \frac{1}{n} \Sigma_{SS}^{-1} X_S^\top (X_S \beta^*_S + \Pi_S \varepsilon - \Pi_S X_{S^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})) \\ &= \beta^*_S + \frac{1}{n} \Sigma_{SS}^{-1} X_S^\top \varepsilon - \Sigma_{SS}^{-1} \Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c}). \end{aligned} \tag{F.4}$$

It remains to control the two terms $A_S = \frac{1}{n} \Sigma_{SS}^{-1} X_S^\top \varepsilon$ and $\Sigma_{SS}^{-1} \Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})$. For the second term, we have

$$\begin{aligned} \|\Sigma_{SS}^{-1} \Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})\|_\infty &\leq \max_{\|v\|_\infty = 1} \|\Sigma_{SS}^{-1} v\|_\infty \|\Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})\|_\infty \\ &\overset{(4.5)}{\leq} K(S)\, (\|\widehat{\beta}^{(P1)}\|_1 + \|\beta^*_{S^c}\|_1). \end{aligned} \tag{F.5}$$

Step 3: *Putting together the pieces.* The two random terms A and A_S are maxima of a finite collection of linear combinations of sub-Gaussian random variables so that (A.2) in Appendix A can be applied by estimating Euclidean norms. For A, we use (F.2). Second, we have

$$A_S = \max_{1 \leq j \leq s} \frac{|v_j^\top \varepsilon|}{n}, \quad v_j = X_S \Sigma_{SS}^{-1} e_j, \;\; j = 1, \ldots, s, \tag{F.6}$$

where e_j denotes the j-th canonical basis vector. One has

$$\max_{1 \leq j \leq s} \|v_j\|_2^2 = \max_{1 \leq j \leq s} e_j^\top \Sigma_{SS}^{-1} X_S^\top X_S \Sigma_{SS}^{-1} e_j \overset{(4.5)}{\leq} \frac{n}{\phi_{\min}(S)}.$$

It follows that for any $M \geq 0$ the event

$$\left\{ A \leq 2(1+M)\sigma\sqrt{\frac{2\log p}{n}} \right\} \cap \left\{ A_S \leq \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(S)}}\sqrt{\frac{2\log p}{n}} \right\}$$

holds with probability no less than $1 - 4p^{-M^2}$. Conditional on that event, it follows that with b as in Theorem 3, we have

$$\|\beta^*_S - \bar{\beta}^{(P2)}\|_\infty \leq (b + \|\beta^*_{S^c}\|_1)K(S) + \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(S)}}\sqrt{\frac{2\log p}{n}},$$

and hence, using the lower bound on $\beta_{\min}(S)$, that $\bar{\beta}^{(P2)} = \widehat{\beta}_S \succ 0$ and thus also that $\widehat{\beta}^{(P1)} = \widehat{\beta}_{S^c}$. □

Appendix G: Proof of Theorem 4

Part 1: proof of uniqueness. To prove the first part of the theorem asserting uniqueness of the NNLS solution, we need two additional lemmas. The first one is a concentration result which is a special case of Theorem 2.5 in [28].

Lemma 4. *Let $\Pi \in \mathbb{R}^{n\times n}$ be a projection matrix on a d-dimensional subspace of $\mathbb{R}^n$ and let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ be a random vector whose entries are i.i.d. zero-mean sub-Gaussian random variables with parameter σ. Then*

$$\mathbf{P}\left(\|\Pi\varepsilon\|_2^2 \leq \mathbf{E}[\varepsilon_1^2]\,\frac{d}{4} \right) \leq 2\exp\left(-\frac{c}{\sigma^4}d\right),$$

where $c > 0$ is a universal constant.

The second lemma provides two sufficient conditions for the NNLS solution to be unique.

Lemma 5. *Let the columns of X be in general linear position. Then the NNLS problem has a unique solution if one of the following holds:*

(i) $p \leq n$, *(ii)* $\min_{\beta\succeq 0} \frac{1}{n}\|y - X\beta\|_2^2 > 0$.

Moreover, under (ii) the active set $F = \{j : \widehat{\beta}_j > 0\}$ satisfies $|F| \leq \min\{n-1, p\}$. Conversely, if y has a distribution that is absolutely continuous with respect to the Lebesgue measure, then $|F| \leq \min\{n-1, p\}$ implies with probability one that the NNLS problem has a unique solution.

Proof. Suppose that (i) holds. The fact that the columns of X are in general linear position implies that Σ is strictly positive definite so that the NNLS objective is strictly convex and hence has a unique minimizer. We now turn to the case $p > n$. We first note that $X\widehat{\beta}$ is unique, because it is the projection of y onto the polyhedral cone $\mathcal{C} = \{z \in \mathbb{R}^n : z = X\beta,\ \beta \in \mathbb{R}^p_+\}$, which is a convex set. Moreover, under (ii), $X\widehat{\beta}$ must be contained in the boundary $\partial\mathcal{C}$ of $\mathcal{C}$ (by

general linear position, the interior of $\mathcal{C}$ is non-empty). Note that $\partial\mathcal{C}$ equals the union of the facets of $\mathcal{C}$, that is

$$\partial\mathcal{C} = \bigcup_{J\in\mathcal{F}} \mathcal{C}_J, \quad \text{where } \mathcal{C}_J = \{z \in \mathcal{C} : z = X_J\beta,\ \beta \in \mathbb{R}^{n-1}_+\}$$
$$\text{and } \mathcal{F} = \{J \subset \{1,\dots,p\},\ |J| = n-1 : \exists w \in \mathbb{R}^n \text{ s.t. } z^\top w = 0\ \forall z \in \mathcal{C}_J \text{ and } z^\top w > 0\ \forall z \in \mathcal{C}\setminus\mathcal{C}_J\}.$$

From $X\widehat{\beta} \in \partial\mathcal{C}$, it hence follows that the active set $F = \{j : \widehat{\beta}_j > 0\}$ has cardinality at most $n-1$. General linear position implies that the linear system $X_F\beta = X\widehat{\beta}$ has exactly one solution $\beta = \widehat{\beta}_F$.

Concerning the second part of the lemma, the assertion $|F| \le \min\{n-1, p\}$ is trivial for $p \le (n-1)$. Conversely, if $p \ge n$, the fact that $\min_{\beta\succeq 0} \frac{1}{n}\|y - X\beta\|_2^2 > 0$ allows us to conclude that $X\widehat{\beta} \in \partial\mathcal{C}$ so that the assertion follows from the reasoning above. For the third part, we note that the fact that y has a distribution which is absolutely continuous with respect to the Lebesgue measure implies that y is not contained in any subspace of dimension smaller than n with probability one, so that $\min_{\beta\succeq 0} \frac{1}{n}\|y - X\beta\|_2^2 > 0$, and the claim follows from part (ii). □

Using Lemma 5 and the condition (4.7)

$$\frac{32(1+M)^2\sigma^2}{\mathbf{E}[\varepsilon_1^2]}\ \frac{\log p}{\tau^2(S)\, n} \le \left(1 - \frac{s}{n}\right),$$

we will show that for $p \ge n$, condition (ii) of Lemma 5 holds with the stated probability, from which we will conclude the proof of the first part of the theorem. Note that for $p \le n-1$, uniqueness follows from general linear position while the claim $|F| \le \min\{n-1, p\}$ is trivial. Let us recall the decomposition of Lemma 2. Note that

$$\min_{\beta\succeq 0} \frac{1}{n}\|y - X\beta\|_2^2 \ge \min_{\beta^{(P1)}\succeq 0} \frac{1}{n}\|\xi - Z\beta^{(P1)}\|_2^2,$$

hence it suffices to show that the right hand side is strictly positive. Suppose conversely that $\xi = Z\widehat{\beta}^{(P1)}$, then $\frac{1}{n}\|\xi\|_2^2 = \frac{1}{n}\|Z\widehat{\beta}^{(P1)}\|_2^2$. Since $\widehat{\beta}^{(P1)}$ is a minimizer of $(P1)$, $\frac{1}{n}\|Z\widehat{\beta}^{(P1)}\|_2^2 \le \frac{2}{n}\xi^\top Z\widehat{\beta}^{(P1)}$, which, by the definition of $\tau^2(S)$, implies that

$$\|\widehat{\beta}^{(P1)}\|_1 \le \frac{1}{\tau^2(S)}\frac{2}{n}\|Z^\top\xi\|_\infty$$

and in turn

$$\frac{1}{n}\|\Pi_S^\perp\varepsilon\|_2^2 = \frac{1}{n}\|\xi\|_2^2 = \frac{1}{n}\|Z\widehat{\beta}^{(P1)}\|_2^2 \le \frac{1}{\tau^2(S)}\left(\frac{2}{n}\|Z^\top\xi\|_\infty\right)^2$$

Hence, conditional on the event

$$\left\{\|\Pi_S^\perp\varepsilon\|_2^2 > \mathbf{E}[\varepsilon_1^2]\,\frac{n-s}{4}\right\} \cap \left\{\left(\frac{2}{n}\|Z^\top\xi\|_\infty\right)^2 \le 8(1+M)^2\sigma^2\,\frac{\log p}{n}\right\} \quad \text{(G.1)}$$

it holds that

$$\frac{\mathbf{E}[\varepsilon_1^2]}{4}\left(1-\frac{s}{n}\right) < \frac{1}{n}\|\xi\|_2^2 \leq 8(1+M)^2\sigma^2\frac{\log p}{\tau^2(S)\,n},$$

which contradicts (4.7). As a result, $\min_{\beta\succeq 0}\frac{1}{n}\|y-X\beta\|_2^2 > 0$ as was to be shown. Invoking Lemma 4 with $\Pi = \Pi_S^{\perp}$ so that $d = n-s$ by general linear position and treating the second event in (G.1) as in step 3 of Appendix F, the probability of the event (G.1) is no less than $1-\exp(-c(n-s)/\sigma^4) - 2p^{-M^2}$.

Part 2: proof of the bound on $\|\widehat{\beta}-\beta^*\|_\infty$. Given uniqueness of the NNLS solution and in turn of its active set $F = \{j : \widehat{\beta}_j > 0\}$, the stated bound on $\|\widehat{\beta}-\beta^*\|_\infty$ follows readily once it holds that $S \subseteq F$. In fact, the optimality conditions of the NNLS problem (cf. Lemma 3) then yield that $\widehat{\beta}_F$ can be recovered from the linear system

$$\Sigma_{FF}\widehat{\beta}_F = \frac{X_F^\top(X_S\beta_S^* + \varepsilon)}{n} = \frac{X_F^\top(X_F\beta_F^* + \varepsilon)}{n},$$

where the second equality results from $S \subseteq F$. As an immediate consequence, we have that

$$\|\widehat{\beta}-\beta^*\|_\infty = \|\widehat{\beta}_F - \beta_F^*\|_\infty = \|\Sigma_{FF}^{-1}X_F^\top\varepsilon/n\|_\infty.$$

In order to control the random term, we may follow the reasoning below (F.6) to conclude that for any $M \geq 0$, the event

$$\{\|\Sigma_{FF}^{-1}X_F^\top\varepsilon/n\|_\infty \leq (1+M)\sigma\,\{\phi_{\min}(F)\}^{-1/2}\sqrt{2\log(p)/n}\}$$

has probability at least $1-2p^{-M^2}$. It remains to show that under the conditions of the theorem, we indeed have that $S \subseteq F$. This is done by referring to the scheme in Appendix F . Given the lower bound on $\beta_{\min}(S)$, it is established that the event $\{\widehat{\beta}_S = \widehat{\beta}^{(P2)} \succ 0\}$ and in turn $\{S \subseteq F\}$ occurs with probability at least $1-4p^{-M^2}$. This finishes the proof.

Appendix H: Proof of Theorem 5

We first recall that the analysis is conditional on the event

$$E = \{r_j \leq s \text{ for all } j \in S\}, \quad \text{where } r_j = k \Leftrightarrow \widehat{\beta}_j = \widehat{\beta}_{(k)}. \tag{H.1}$$

Our proof closely follows the corresponding proof in [23]. We show in two steps that both $S \setminus \widehat{S} = \emptyset$ and $\widehat{S} \setminus S = \emptyset$. For both steps, we shall need the following observations. Let V_k denote the linear space spanned by the top k variables according to the given ranking, $k = 1,\ldots,p$, and let $V_0 = \{0\}$. Let further $U_k = V_k^{\perp} \cap V_{k+1}$, $k = 0,\ldots,p-1$, which are subspaces of $\mathbb{R}^n$ of dimension at most 1. In case that the dimension of U_k is one, let u_k be the unit vector

spanning U_k and let $u_k = 0$ otherwise, $k = 0, \ldots, p-1$. Note that $\Pi(k+1) - \Pi(k)$ as appearing in the definition of the $\delta(k)$'s equals the projection on the U_k, $k = 0, \ldots, p-1$. In particular, we have

$$\|(\Pi(k+1) - \Pi(k))\varepsilon\|_2 = |\langle u_k, \varepsilon \rangle|, \quad k = 0, \ldots, p-1. \tag{H.2}$$

Step 1: *No false negatives.* In the sequel, let Δ denote the threshold of the procedure so that

$$\widehat{s} = \max\{0 \leq k \leq (p-1) : \delta(k) \geq \Delta\} + 1.$$

Later in the proof, it will be verified that Δ can be chosen as asserted in the theorem. We first note that conditional on E, by definition of $\widehat{s}$, it holds that the event $\{S \setminus \widehat{S} = \emptyset\}$ is contained in the event $\{\delta(s-1) \geq \Delta\}$. Hence it suffices to upper bound the probability of the event $\{\delta(s-1) < \Delta\}$. We have

$$\begin{aligned}
\mathbf{P}(\delta(s-1) < \Delta) &= \mathbf{P}\left(\|(\Pi(s) - \Pi(s-1))y\|_2 < \Delta\right) \\
&\leq \mathbf{P}\left(\|(\Pi(s) - \Pi(s-1))X_S\beta^*_S\|_2 < \Delta + \|(\Pi(s) - \Pi(s-1))\varepsilon\|_2\right) \\
&\overset{(H.2)}{=} \mathbf{P}\left(\|(\Pi(s) - \Pi(s-1))X_S\beta^*_S\|_2 < \Delta + |\langle u_{s-1}, \varepsilon \rangle|\right) \\
&\leq \mathbf{P}\left(\min_{j \in S}\|(\Pi_S - \Pi_{S\setminus j})X_j\beta^*_j\|_2 < \Delta + |\langle u_{s-1}, \varepsilon \rangle|\right),
\end{aligned} \tag{H.3}$$

where Π_S and $\Pi_{S\setminus j}$ denote the projection on the linear spaces spanned by the columns of X corresponding to S respectively $S \setminus j$, $j = 1, \ldots, s$. In order to obtain the second inequality, we have used again that we work conditional on the event E. As will be established at the end of the proof, we further have

$$\min_{j \in S}\|(\Pi_S - \Pi_{S\setminus j})X_j\beta^*_j\|_2 \geq \sqrt{n}\left\{\phi_{\min}(S)\right\}^{1/2}\beta_{\min}(S). \tag{H.4}$$

Combining (H.3) and (H.4), it suffices to upper bound

$$\mathbf{P}\left(|\langle u_{s-1}, \varepsilon \rangle| > \sqrt{n}\left\{\phi_{\min}(S)\right\}^{1/2}\beta_{\min}(S) - \Delta\right) \tag{H.5}$$

as will be done below after fixing Δ.

Step 2: *No false positives.* Conditional on E, the probability of having a false positive selection is upper bounded as

$$\begin{aligned}
\mathbf{P}(\cup_{k=s}^{p-1}\{\delta(k) \geq \Delta\}) &= \mathbf{P}\left(\max_{s \leq k \leq p-1}\|(\Pi(k+1) - \Pi(k))y\|_2 \geq \Delta\right) \\
&= \mathbf{P}\left(\max_{s \leq k \leq p-1}\|(\Pi(k+1) - \Pi(k))\varepsilon\|_2 \geq \Delta\right) \\
&= \mathbf{P}\left(\max_{s \leq k \leq p-1}|\langle u_k, \varepsilon \rangle| \geq \Delta\right).
\end{aligned} \tag{H.6}$$

Choosing $\Delta = (1+M)\sigma\sqrt{2\log(p)}$ for an arbitrary $M \geq 0$, using the assumption on $\beta_{\min}(S)$, and controlling $\max_{0\leq k\leq p-1} |\langle u_k, \varepsilon\rangle|$ according to (A.2) in the usual way, the probabilities (H.5) and (H.6) do not exceed $2p^{-M^2}$. The assertion of the theorem then follows. To conclude the proof, it remains to establish (H.4). Let us fix an arbitrary $j \in S$. We have

$$\|(\Pi_S - \Pi_{S\setminus j})X_j\|_2 = \|X_j - \Pi_{S\setminus j}X_j\|_2 = \sqrt{\|X_j\|_2^2 - \|\Pi_{S\setminus j}X_j\|_2^2}$$

Write θ for the vector of regression coefficients for the linear regression of X_j on $\{X_k\}_{k\in S\setminus\{j\}}$, i.e.

$$\theta = (X_{S\setminus j}^\top X_{S\setminus j})^{-1}X_{S\setminus j}^\top X_j,$$

and note that, according to a block decomposition of the matrix $X_S^\top X_S$

$$\begin{aligned}\begin{pmatrix}-\theta\\1\end{pmatrix}^\top (X_S^\top X_S)\begin{pmatrix}-\theta\\1\end{pmatrix} &= \begin{pmatrix}-\theta\\1\end{pmatrix}^\top \begin{pmatrix}X_{S\setminus j}^\top X_{S\setminus j} & X_{S\setminus j}^\top X_j\\ X_j^\top X_{S\setminus j} & \|X_j\|_2^2\end{pmatrix}\begin{pmatrix}-\theta\\1\end{pmatrix}\\ &= \|X_j\|_2^2 - X_j^\top X_{S\setminus j}(X_{S\setminus j}^\top X_{S\setminus j})^{-1}X_{S\setminus j}^\top X_j\\ &= \|X_j\|_2^2 - \|\Pi_{S\setminus j}X_j\|_2^2.\end{aligned}$$

We conclude the proof from $X_S^\top X_S = n\Sigma_{SS}$ and

$$\begin{pmatrix}-\theta\\1\end{pmatrix}^\top (X_S^\top X_S)\begin{pmatrix}-\theta\\1\end{pmatrix} \geq n\phi_{\min}(S)\left\|\begin{pmatrix}-\theta\\1\end{pmatrix}\right\|_2^2 \geq n\phi_{\min}(S).$$

Appendix I: Proof of Theorem 6

Consider the non-negative lasso problem (1.3). It follows from the KKT optimality conditions that any minimizer $\widehat{\beta}^{\ell_{1,\succeq}}$ of (1.3) satisfies

$$\begin{aligned}&\frac{2}{n}X_j^\top(y - X\widehat{\beta}^{\ell_{1,\succeq}}) = \lambda \ \text{ and } \ \widehat{\beta}_j^{\ell_{1,\succeq}} > 0,\\ \text{or}\quad &\frac{2}{n}X_j^\top(y - X\widehat{\beta}^{\ell_{1,\succeq}}) \leq \lambda \ \text{ and } \ \widehat{\beta}_j^{\ell_{1,\succeq}} = 0, \quad j = 1,\ldots,p.\end{aligned} \tag{I.1}$$

Following the technique employed in [54], we establish that under the conditions of the theorem, the unique minimizer of the non-negative lasso problem is given by $\widehat{\beta}_S^{\ell_{1,\succeq}} = \widehat{\alpha}_S \succ 0$ and $\widehat{\beta}_{S^c}^{\ell_{1,\succeq}} = 0$ with the specified probability, where $\widehat{\alpha}$ denotes the minimizer of the following constrained non-negative lasso problem

$$\min_{\beta_S\succeq 0,\ \beta_{S^c}=0} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\mathbf{1}^\top\beta. \tag{I.2}$$

To this end, in view of (I.1), it suffices to show that the following system of inequalities is satisfied

$$\frac{2}{n}\begin{bmatrix}X_S^\top\varepsilon\\ X_{S^c}^\top\varepsilon\end{bmatrix} + 2\begin{bmatrix}\Sigma_{SS} & \Sigma_{SS^c}\\ \Sigma_{S^cS} & \Sigma_{S^cS^c}\end{bmatrix}\begin{bmatrix}\beta_S^* - \widehat{\alpha}_S\\ 0\end{bmatrix}\begin{bmatrix}=\\ \prec\end{bmatrix}\begin{bmatrix}\lambda\mathbf{1}\\ \lambda\mathbf{1}\end{bmatrix}. \tag{I.3}$$

In view of the required lower bound on $\beta_{\min}(S)$ in (4.14), we have

$$0 \prec \widehat{\alpha}_S = \beta^*_S - \frac{\lambda}{2}\Sigma_{SS}^{-1}\mathbf{1} + \Sigma_{SS}^{-1}\frac{1}{n}X_S^\top \varepsilon \tag{I.4}$$

with probability at least $1 - 2p^{-M^2}$, handling the random term as (F.6) in the proof of Theorem 3. Substituting (I.4) back into (I.3), we find that the following system of inequalities must hold true:

$$\frac{\lambda}{2}\left\{\Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbf{1}\right\} + \frac{1}{n}X_{S^c}^\top(I - \Pi_S)\varepsilon \prec \frac{\lambda}{2}\mathbf{1}. \tag{I.5}$$

In light of the maximal inequality (A.2), the event

$$\left\{\max_{j\in S^c}\frac{1}{n}X_j^\top(I-\Pi_S)\varepsilon \leq \lambda_M\right\}, \quad \lambda_M = \sigma(1+M)\sqrt{2\log(p)/n}, \tag{I.6}$$

occurs with probability at least $1 - 2p^{-M^2}$, noting that $\|(I-\Pi_S)X_j\|_2 \leq \|X_j\|_2$ for all $j = 1, \ldots, p$. Hence, conditional on the events $\{\widehat{\alpha}_S \succ 0\}$ and (I.6), each component of the left hand side of (I.5) is no larger than $\frac{\lambda}{2}\iota(S) + \lambda_M$ (cf. definition (4.12)), so that for $\lambda > 2\lambda_M/(1-\iota(S))$, the system of inequalities (I.5) and hence also (I.3) are indeed fulfilled.

Appendix J: Example of a design for which the non-negative lasso performs sub-optimally

As indicated in the discussion following Theorem 6, the non-negative lasso does not always attain the optimal rate for estimating β^* with respect to the ℓ_∞-norm. For the data-generating model of Theorem 6 we give an example of a design so that the non-negative lasso estimator has the property

$$\|\widehat{\beta}^{\ell_{1,\geq}} - \beta^*\|_\infty = \Omega\left(\sqrt{s\log(p)/n}\right), \tag{J.1}$$

with high probability, provided the regularization parameter $\lambda = \Omega(\sqrt{\log(p)/n})$ (as conventionally suggested in the literature) and the minimum support coefficient $\beta_{\min}(S) = \Omega(\sqrt{s\log(p)/n})$. For the same design, if $\beta_{\min}(S) = \Omega(\sqrt{\log(p)/n})$, the NNLS estimator obeys the bound

$$\|\widehat{\beta} - \beta^*\|_\infty = O\left(\sqrt{\log(p)/n}\right), \tag{J.2}$$

with high probability. In order to establish (J.1), we shall build on the scheme used for the proof of Theorem 6 in the previous paragraph. Consider a design whose Gram matrix is of the form

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & 0 \\ 0 & I_{p-s} \end{bmatrix},$$

where

$$\Sigma_{SS} = \begin{bmatrix} 1 & -1/\sqrt{2(s-1)}\mathbf{1}_{s-1}^\top \\ -1/\sqrt{2(s-1)}\mathbf{1}_{s-1} & I_{s-1} \end{bmatrix}.$$

The constant $\sqrt{2}$ in the denominator is chosen for convenience; any other constant larger than 1 would do as well. Using Schur complements, one computes that

$$\Sigma_{SS}^{-1} = \begin{bmatrix} 2 & \sqrt{2/(s-1)}\mathbf{1}_{s-1}^\top \\ \sqrt{2/(s-1)}\mathbf{1}_{s-1} & I_{s-1} + \frac{1}{s-1}\mathbf{1}_{s-1}\mathbf{1}_{s-1}^\top \end{bmatrix}.$$

As a result, we have that

$$e_1^\top \Sigma_{SS}^{-1}\mathbf{1} = 2 + \sqrt{2(s-1)} = \Omega(\sqrt{s}), \tag{J.3}$$

where e_1 is the first canonical basis vector. Furthermore, the sequence of eigenvalues of Σ_{SS} (ordered decreasingly) is given by

$$\phi_1 = 1 + \frac{1}{\sqrt{2}}, \quad \phi_2 = \cdots = \phi_{s-1} = 1, \quad \phi_s = 1 - \frac{1}{\sqrt{2}}.$$

From the proof of Theorem 6, we know that given the active set $Q = \{j : \widehat{\beta}_j^{\ell_{1,\succeq}} > 0\}$, the non-negative lasso estimator has the following closed form expression.

$$\widehat{\beta}_Q^{\ell_{1,\succeq}} = \beta_Q^* - \frac{\lambda}{2}\Sigma_{QQ}^{-1}\mathbf{1} + \Sigma_{QQ}^{-1}\frac{1}{n}X_Q^\top\varepsilon, \quad \text{and} \quad \widehat{\beta}_{Q^c}^{\ell_{1,\succeq}} = 0. \tag{J.4}$$

If $S \nsubseteq Q$, the claim (J.1) follows trivially from the assumption $\beta_{\min}(S) = \Omega(\sqrt{s\log(p)/n})$. Conversely, if $S \subseteq Q$, (J.4) and the block structure of Σ imply that

$$\begin{aligned}
\|\widehat{\beta}^{\ell_{1,\succeq}} - \beta^*\|_\infty &\geq \|\widehat{\beta}_Q^{\ell_{1,\succeq}} - \beta_Q^*\|_\infty \\
&\geq \|\widehat{\beta}_S^{\ell_{1,\succeq}} - \beta_S^*\|_\infty \\
&\geq \frac{\lambda}{2}e_1^\top\Sigma_{SS}^{-1}\mathbf{1} - \left|e_1^\top\Sigma_{SS}^{-1}\frac{1}{n}X_S^\top\varepsilon\right| \\
&\geq \frac{\lambda}{2}e_1^\top\Sigma_{SS}^{-1}\mathbf{1} - \frac{1}{\sqrt{\phi_s}}\max_{1\leq j\leq p}\left|\frac{X_j^\top\varepsilon}{n}\right| = \Omega\left(\sqrt{s\log(p)/n}\right),
\end{aligned}$$

using (J.3) and the maximal inequality (A.2) to upper bound the second term after the third inequality. The latter scales as $O(\sqrt{\log(p)/n})$ with probability at least $1 - O((p \vee n)^{-1})$. Furthermore, we have used that $\lambda = \Omega(\sqrt{\log(p)/n})$ and $\phi_s = \Omega(1)$.

Regarding the upper bound (J.2) for NNLS, we note that the optimality conditions of the NNLS problem (Lemma 3) in conjunction with the block structure of Σ imply that

$$\widehat{\beta}_S \succ 0 \quad \text{and} \quad \|\widehat{\beta}_S - \beta_S^*\|_\infty \leq \|\Sigma_{SS}^{-1}X_S^\top\varepsilon/n\|_\infty \leq \frac{1}{\sqrt{\phi_s}}\|X_S^\top\varepsilon/n\|_\infty = O(\sqrt{\log(p)/n}),$$

with probability at least $1 - O((p \vee n)^{-1})$, provided $\beta_{\min}(S)$ exceeds the bound on the right hand side. Similarly,

$$\|\widehat{\beta}_{S^c}\|_\infty \leq \|X_{S^c}^\top \varepsilon/n\|_\infty = O(\sqrt{\log(p)/n})$$

with probability at least $1 - O((p \vee n)^{-1})$, so that (J.2) follows.

Appendix K: Proof of Proposition 2

We start by noting that Σ is strictly positive definite so that the NNLS problem is strictly convex. Thus, the NNLS solution and its active set $F = \{j : \widehat{\beta}_j > 0\}$ are unique. Let us first consider the case $s > 0$. Using a slight modification of the scheme used in the proofs of Theorems 3 and 4, we will show that under the required condition on $\beta_{\min}(S)$, the event $\{\widehat{\beta}_S = \widehat{\beta}^{(P2)} \succ 0\}$ holds with the stated probability, which proves the first statement of the proposition. Following the proof of Theorem 3, we have that

$$\|\widehat{\beta}^{(P1)}\|_1 \leq \frac{2(1+M)\sigma\sqrt{2\log(p)/n}}{\tau^2(S)} \overset{(5.5)}{\leq} \frac{2(1+(s-1)\rho)(1+M)\sigma\sqrt{2\log(p)/n}}{\rho(1-\rho)},$$

with probability at least $1 - 2p^{-M^2}$, where we have used the closed form expression for $\tau^2(S)$ in (5.5). In order to verify that $\widehat{\beta}_S = \widehat{\beta}^{(P2)} \succ 0$, we follow the back-substitution step (step 2 in Appendix F) apart from the following modification. In place of (F.5), we bound

$$\begin{aligned}\|\Sigma_{SS}^{-1}\Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})\|_\infty &= \rho\|\Sigma_{SS}^{-1}\mathbf{1}\|_\infty\|\widehat{\beta}^{(P1)}\|_1 \\ &\leq \frac{\rho}{1+(s-1)\rho}\|\widehat{\beta}^{(P1)}\|_1 \leq \frac{2(1+M)\sigma\sqrt{2\log(p)/n}}{1-\rho}\end{aligned}$$

For the first equality, we have used that $\beta^*_{S^c} = 0$ and the fact that the matrix Σ_{SS^c} has constant entries equal to ρ. For the second inequality, we have used that $\mathbf{1}$ is an eigenvector of Σ_{SS} corresponding to its largest eigenvalue $1+(s-1)\rho$. Turning to step 3 in Appendix F, we note that with $\phi_{\min}(S) = (1-\rho)$,

$$\begin{aligned}\|\beta^*_S - \bar{\beta}^{(P2)}\|_\infty &\leq \frac{2(1+M)\sigma\sqrt{2\log(p)/n}}{1-\rho} + \frac{(1+M)\sigma\sqrt{2\log(p)/n}}{\sqrt{1-\rho}} \\ &\leq \frac{3(1+M)\sigma\sqrt{2\log(p)/n}}{1-\rho}\end{aligned}$$

so that $\bar{\beta}^{(P2)} = \widehat{\beta}^{(P2)} = \widehat{\beta}_S \succ 0$ with probability at least $1 - 4p^{-M^2}$ as claimed.

We now turn to the second statement of the proposition concerning the (conditional) distribution of the cardinality of the active set. Conditional on the event $\{\widehat{\beta}_S \succ 0\}$, the KKT optimality conditions of the NNLS problem as stated in Lemma 3 imply that the following block system of inequalities holds.

$$\begin{bmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^cS} & \Sigma_{S^cS^c} \end{bmatrix} \begin{bmatrix} \widehat{\beta}_S \\ \widehat{\beta}_{S^c} \end{bmatrix} \begin{bmatrix} = \\ \preceq \end{bmatrix} \begin{bmatrix} \Sigma_{SS}\beta^*_S + \frac{X_S^\top \varepsilon}{n} \\ \Sigma_{S^cS}\beta^*_S + \frac{X_{S^c}^\top \varepsilon}{n} \end{bmatrix}. \tag{K.1}$$

Resolving the top block for $\widehat{\beta}_S$, we obtain

$$\widehat{\beta}_S = \beta^*_S + \Sigma_{SS}^{-1}\left(\frac{X_S^\top \varepsilon}{n} - \Sigma_{SS^c}\widehat{\beta}_{S^c}\right).$$

Back-substituting that expression into the bottom block of inequalities yields the following system of inequalities.

$$\left(\Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}\right)\widehat{\beta}_{S^c} \preceq \frac{X_{S^c}^\top (I - X_S(X_S^\top X_S)^{-1}X_S^\top)\varepsilon}{n} = \frac{Z^\top \varepsilon}{n}, \quad \text{(K.2)}$$

where $Z = \Pi_S^\perp X_{S^c}$. For equi-correlated design with $\Sigma = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$, we have that

$$\Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c} = (1-\rho)I + \underbrace{\frac{\rho(1-\rho)}{1+(s-1)\rho}}_{\gamma(s,\rho)}\mathbf{1}\mathbf{1}^\top = (1-\rho)I + \gamma(s,\rho)\mathbf{1}\mathbf{1}^\top, \quad \text{(K.3)}$$

cf. the derivation in (5.5). Denote $\widehat{\alpha} = \widehat{\beta}_{S^c}$, and $G = \{k : \ \widehat{\alpha}_k > 0\}$. Using Lemma 3 and (K.3), (K.2) can be written as

$$\begin{aligned} \frac{Z_k^\top \varepsilon}{n} - (1-\rho)\widehat{\alpha}_k &= \gamma(s,\rho)\,\mathbf{1}^\top \widehat{\alpha}, \quad k \in G, \\ \frac{Z_k^\top \varepsilon}{n} &\le \gamma(s,\rho)\mathbf{1}^\top \widehat{\alpha}, \quad k \notin G. \end{aligned} \quad \text{(K.4)}$$

Set $z = Z^\top \varepsilon/(\sigma\sqrt{n})$ so that z is a zero-mean Gaussian random vector with covariance

$$\frac{1}{n}Z^\top Z = \frac{1}{n}X_{S^c}^\top \Pi_S^\perp X_{S^c} = \Sigma_{S^cS^c} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}.$$

In view of (K.3), z has the distribution as claimed in Proposition 2. From (K.4), we conclude that

$$k \in G \Rightarrow z_k > 0 \ \text{ and } \ z_k \le z_l \text{ for } l \notin G \Rightarrow k \notin G. \quad \text{(K.5)}$$

In particular, recalling that $z_{(1)} \ge z_{(2)} \ge \cdots \ge z_{(p-s)}$ denotes the arrangement of the components of z in decreasing order, if $z_{(1)} \le 0$, then (K.4) implies that $\widehat{\alpha} = 0$, $G = \emptyset$ and $|F| = s$ as stated in the proposition. Let us henceforth assume that $z_{(1)} > 0$, in which case (K.4) implies that G is non-empty. We may then resolve the first set of equations in (K.4) with respect to $\widehat{\alpha}_G$, which yields

$$\widehat{\alpha}_G = \left((1-\rho)I + \gamma(s,\rho)\mathbf{1}\mathbf{1}^\top\right)^{-1}\frac{Z_G^\top \varepsilon}{n} = \frac{1}{1-\rho}\left(\frac{Z_G^\top \varepsilon}{n} - \frac{\gamma(s,\rho)\mathbf{1}\mathbf{1}^\top (Z_G^\top \varepsilon/n)}{(1-\rho)+\gamma(s,\rho)|G|}\right),$$

where the second equality is an application of the Sherman-Woodbury-Morrison formula. This implies in turn that

$$\mathbf{1}^\top \widehat{\alpha} = \mathbf{1}^\top \widehat{\alpha}_G = \frac{\mathbf{1}^\top Z_G^\top \varepsilon}{n}\frac{1}{(1-\rho)+|G|\gamma(s,\rho)}.$$

Substituting this expression back into (K.4), we obtain

$$\begin{aligned}
&\frac{\frac{Z_k^\top \varepsilon}{n}}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} - \widehat{\alpha}_k \frac{(1-\rho) + |G|\gamma(s,\rho)}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} = \frac{\gamma(s,\rho)}{1-\rho}, \quad k \in G, \\
&\frac{\frac{Z_k^\top \varepsilon}{n}}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} \leq \frac{\gamma(s,\rho)}{1-\rho}, \quad k \notin G.
\end{aligned} \tag{K.6}$$

Now note that for $k = 1, \ldots, p-s$,

$$\frac{\frac{Z_k^\top \varepsilon}{n}}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} = \frac{\frac{Z_k^\top \varepsilon}{\sqrt{n}}}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{\sqrt{n}} - \frac{Z_k^\top \varepsilon}{\sqrt{n}}\right)} = \frac{z_k}{\sum_{\ell \in G}(z_\ell - z_k)}$$

From

$$\underbrace{\frac{z_{(2)}}{(z_{(1)} - z_{(2)})}}_{\zeta_1} \geq \underbrace{\frac{z_{(3)}}{(z_{(1)} - z_{(3)}) + (z_{(2)} - z_{(3)})}}_{\zeta_2} \geq \cdots \geq \underbrace{\frac{z_{(p-s)}}{\sum_{k=1}^{p-s-1}(z_{(k)} - z_{(p-s)})}}_{\zeta_{p-s-1}},$$

(K.5) and the inequalities in (K.6), it then follows that

$$\begin{aligned}
G &= \{j : z_j = z_{(1)}\} \cup \left\{k \neq j : \frac{z_k}{\sum_{\ell : z_\ell \geq z_k}(z_\ell - z_k)} > \frac{\gamma(s,\rho)}{1-\rho}\right\} \\
&= \{j : z_j = z_{(1)}\} \cup \left\{k \neq j : \frac{z_k}{\sum_{\ell : z_\ell \geq z_k}(z_\ell - z_k)} \geq \zeta_m\right\},
\end{aligned}$$

where m is the largest integer so that $\zeta_m > \gamma(s,\rho)/(1-\rho) = \theta(s,\rho)$ with $\theta(s,\rho)$ as defined in (5.6), which finishes the proof for $s > 0$. Turning to the case $s = 0$, a similar scheme can be used, starting from the system of inequalities $\Sigma\widehat{\beta} \preceq \frac{X^\top \varepsilon}{n} = \frac{Z^\top \varepsilon}{n} = \sigma z/\sqrt{n}$. The expressions used above remain valid with $\gamma(0,\rho) = \rho$.

Appendix L: Proofs of Propositions 3 and 4

Proofs of Proposition 3 and Proposition 4 are contained in the supplement [45].

References

[1] Szlam, A., Guo, Z., and Osher, S. A split Bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing. In *IEEE International Conference on Image Processing*, 2010.

[2] Bardsley, J. and Nagy, J. Covariance-preconditioned iterative methods for nonnegatively constrained astronomical imaging. *SIAM Journal on Matrix Analysis and Applications*, 27:1184–1198, 2006. MR2205619

[3] Bartlett, P., Mendelson, S., and Neeman, J. ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probability Theory and Related Fields*, 254:193–224, 2012. MR2981422

[4] Belloni, A., Chernozukov, V., and Wang, L. Square root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98:791–806, 2011. MR2860324

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

[6] Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. MR2061575

[7] Bruckstein, A., 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

[8] Bühlmann, P. and van de Geer, S. *Statistics for High-Dimensional Data*. Springer, 2011. MR2807761

[9] Bunea, F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization. *The Electronic Journal of Statistics*, 2:1153–1194, 2008. MR2461898

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

[11] Bunea, F., Tsybakov, A., Wegkamp, M., and Barbu, A. SPADES and mixture models. *The Annals of Statistics*, 38:2525–2558, 2010. MR2676897

[12] Candes, E. and Plan, Y. Near-ideal model selection by ℓ_1-minimization. *The Annals of Statistics*, 37:2145–2177, 2009. MR2543688

[13] Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007. MR2382644

[14] Chen, D. and Plemmons, R. Nonnegativity constraints in numerical analysis. In *Symposium on the Birth of Numerical Analysis*, 2007.

[15] Donoho, D. Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306, 2006

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

[17] Donoho, D. and Tanner, J. Neighbourliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Science*, 102:9452–9457, 2005. MR2168716

[18] Donoho, D. and Tanner, J. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Science*, 102:9446–9451, 2005. MR2168715

[19] Donoho, D. and Tanner, J. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete and Computational Geometry*, 43:522–541, 2010. MR2587835

[20] Dorfman, R. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14:436–440, 1943.

[21] Du, D. and Hwang, F. *Pooling Designs and Nonadaptive Group Testing*. Twayne Publishers, 2006. MR2282446

[22] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004. MR2060166

[23] Genovese, C., Jin, J., Wasserman, L., and Yao, Z. A comparison of the lasso and marginal regression. *Journal of Machine Learning Research*, 13:2107–2143, 2012. MR2956354

[24] Greenshtein, E. and Ritov, Y. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 6:971–988, 2004. MR2108039

[25] Hebiri, M. and Lederer, J. How correlations influence lasso prediction. *IEEE Transactions on Information Theory*, 59:1846–1854, 2013. MR3030757

[26] Huet, S., Giraud, C., and Verzelen, N. High-dimensional regression with unknown variance. *Statistical Science*, 27:500–518, 2012. MR3025131

[27] Kim, D., Sra, S., and Dhillon, I. Tackling box-constrained convex optimization via a new projected quasi-Newton approach. *SIAM Journal on Scientific Computing*, 32:3548–3563, 2010. MR2763737

[28] Latala, R., Mankiewicz, P., Oleskiewicz, K., and Tomczak-Jaegermann, N. Banach-Mazur distances and projections on random subgaussian polytopes. *Discrete and Computational Geometry*, 38:29–50, 2007. MR2322114

[29] Lawson, R. and Hanson, C. *Solving Least Squares Problems*. SIAM Classics in Applied Mathematics, 1987.

[30] Li, L. and Speed, T. Parametric deconvolution of positive spike trains. *The Annals of Statistics*, 28:1279–1301, 2000. MR1805784

[31] Lin, Y., Lee, D., and Saul, L. Nonnegative deconvolution for time of arrival estimation. In *ICASSP*, 2004.

[32] Litvak, A., Pajor, A., Rudelson, M., and Tomczak-Jaegermann, N. Smallest singular value of random matrices and geometry of random polytopes. *Advances in Mathematics*, 195:491–523, 2005. MR2146352

[33] Lounici, K. Sup-norm convergence rate and sign concentration property of lasso and Dantzig selectors. *The Electronic Journal of Statistics*, 2:90–102, 2008. MR2386087

[34] Meinshausen, N. Sign-constrained least squares estimation for high-dimensional regression. *The Electronic Journal of Statistics*, 7:1607–1631, 2013. MR3066380

[35] Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006. MR2278363

[36] Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37:246–270, 2009. MR2488351

[37] Raskutti, G., Wainwright, M., and Yu, B. Restricted nullspace and eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010. MR2719855

[38] Rudelson, M. and Vershynin, R. Non-asymptotic theory of random matrices: Extreme singular values. In *International Congress of Mathematics*, 2010. MR2827856

[39] Rudelson, M. and Zhou, S. Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory*, 59:3434–3447, 2013. MR3061256

[40] Schneider, R. and Weil, W. *Stochastic and Integral Geometry*. Springer, 2008. MR2455326

[41] Schölkopf, B. and Smola, A. *Learning with Kernels*. MIT Press, Cambridge, Massachussets, 2002.

[42] Slawski, M. and Hein, M. Sparse recovery for protein mass spectrometry data. In *NIPS workshop on practical applications of sparse modelling*, 2010.

[43] Slawski, M. and Hein, M. Sparse recovery by thresholded non-negative least squares. In *Advances in Neural Information Processing Systems 24*, pages 1926–1934, 2011.

[44] Slawski, M. and Hein, M. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. arXiv:1205.0953v1, 2012.

[45] Slawski, M. and Hein, M. Supplementary material to 'Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization'. `http://www.ml.uni-saarland.de/people/slawski/nnls_ejs_supplement.pdf`, 2013.

[46] Sun, T. and Zhang, C.-H. Scaled sparse linear regression. *Biometrika*, 99:879–898, 2012. MR2999166

[47] Tibshirani, R. Regression shrinkage and variable selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58:671–686, 1996. MR1379242

[48] Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective (with discussion). *Journal of the Royal Statistical Society Series B*, 73:273–282, 2011. MR2815776

[49] Tropp, J. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50:2231–2242, 2004. MR2097044

[50] Tsybakov, A. Discussion of 'Stability Section' by N. Meinshausen and P. Bühlmann. *Journal of the Royal Statistical Society Series B*, 72:467, 2010.

[51] van de Geer, S. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36:614–645, 2008. MR2396809

[52] VAN DE GEER, S. and BÜHLMANN, P. On the conditions used to prove oracle results for the lasso. *The Electronic Journal of Statistics*, 3:1360–1392, 2009. MR2576316

[53] VAN DE GEER, S. and LEDERER, J. *From Probability to Statistics and Back: High-Dimensional Models and Processes. A Festschrift in Honor of Jon Wellner*, chapter 'The lasso, correlated design, and improved oracle inequalities'. IMS Collections. 2012.

[54] WAINWRIGHT, M. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009. MR2729873

[55] WANG, M. and TANG, A. Conditions for a unique non-negative solution to an underdetermined system. In *Proceedings of Allerton Conference on Communication, Control, and Computing*, 2009.

[56] WANG, M., XU, W., and TANG, A. A unique nonnegative solution to an undetermined system: From vectors to matrices. *IEEE Transactions on Signal Processing*, 59:1007–1016, 2011. MR2797714

[57] WENDEL, J. G. A problem in geometric probability. *Mathematics Scandinavia*, 11:109–111, 1962. MR0146858

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

[59] ZHANG, T. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009. MR2491749

[60] ZHANG, T. Some sharp performance bounds for least squares regression with L_1 regularization. *The Annals of Statistics*, 37:2109–2144, 2009. MR2543687

[61] ZHAO, P. and YU, B. On model selection consistency of the lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006. MR2274449

[62] ZHAO, Y.-B. Sparsest and least ℓ_1-norm nonnegative solutions to linear systems with applications to compressive sensing and linear programs. Technical report, University of Birmingham, 2012.

[63] ZHOU, S. Thresholding procedures for high dimensional variable selection and statistical estimation. In *NIPS*, 2009.

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