UNIVERSITÄT
DES
SAARLANDES

Saarland University
Faculty of Natural Sciences and Technology I
Department of Computer Science

Topics in learning sparse and low-rank models of non-negative data

Martin Slawski

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

Saarbrücken, Juli 2014

Berichterstatter / Reviewers:

Prof. Dr. Matthias Hein
Prof. Dr. Thomas Lengauer, PhD
Prof. Jared Tanner, PhD

Prüfungsausschuss / Examination Board:

Prof. Dr. Joachim Weickert
Prof. Dr. Matthias Hein
Prof. Dr. Thomas Lengauer, PhD
Prof. Jared Tanner, PhD
Dr. Moritz Gerlach

Dekan / Dean:

Prof. Dr. Markus Bläser

Tag des Kolloquiums / Date of the Defense Talk:

February 25th, 2015

Date of this version: April 11th, 2015

Contents

Acknowledgments

First of all, I would like to thank my thesis advisor Matthias Hein for his guidance. I truly admire his outstanding talent to discover interesting directions of research. All of his suggestions have turned out to be rich and rewarding topics. His quickness of mind, efficiency and precision in his work and his ability to get instantly a good grasp of scientific problems with the help of excellent mathematical intuition, as well as his patience with his students (who may not always be able to think as fast as him) have deeply impressed me throughout the work on this thesis. Matthias' skills have contributed to this thesis in many ways, in particular by helping me out several times when I got stuck. At the same, Matthias gave me plenty of freedom to pursue my own ideas. When preparing my work for publication, he always had very good ideas that would improve my presentation, and he saved me quite a few times by spotting subtle (and also less subtle) flaws in my proofs. Lastly, Matthias let me attend several workshops and conferences which not only gave me the possibility to present my own work, but also to learn from the ideas of other researchers.
Next, I would like to thank Prof. Dr. Lengauer, PhD and Prof. Jared Tanner, PhD for acting as reviewers of this thesis. I am also grateful for the discussion I had with Prof. Tanner during the workshop 'Sparsity and Computation' at an early stage of the work on this thesis. In fact, this discussion proved to be rather helpful in later stages.
During the work of this thesis, I was funded by the cluster of excellence 'Multimodel Computing and Interaction' (MMCI) of Deutsche Forschungsgemeinschaft. The financial support is gratefully acknowledged.
I thank all persons with whom I collaborated during this thesis: Barbara Gregorius, Andreas Hildebrandt, Rene Hussong, Thomas Jakoby, David James, Janis Kalofolias, Felix Krahmer, Pavlo Lutsik, Jörn Walter and Qinqing Zheng. Thanks also to Irina Rish and her colleagues for organizing the publication of the book 'Practical Applications of Sparse Modeling'.
I thank all former and present lab members of the machine learning group at Saarland University for a nice working atmosphere.
I also thank my friends for helping me not to ponder about my work all the time. A special thanks goes to Syama Sundar Rangapuram for introducing me to the Indian (especially Telugu-speaking) community in Saarbrücken. Through him, I got to know Srikanth Duddela who has become my best friend here. Without his encouragements, I would probably not have completed this thesis. I will never forgot all the fun we had during table tennis, gym and movie sessions, which helped me to forget about my problems and to stay focused.
This thesis is dedicated to my family. The unconditional support I have experienced throughout my life has been key to my accomplishments.

Abstract

Advances in information and measurement technology have led to a surge in prevalence of high-dimensional data. Sparse and low-rank modeling can both be seen as techniques of dimensionality reduction, which is essential for obtaining compact and interpretable representations of such data.
In this thesis, we investigate aspects of sparse and low-rank modeling in conjunction with non-negative data or non-negativity constraints.
The first part is devoted to the problem of learning sparse non-negative representations, with a focus on how non-negativity can be taken advantage of. We work out a detailed analysis of non-negative least squares regression, showing that under certain conditions sparsity-promoting regularization, the approach advocated paradigmatically over the past years, is not required. Our results have implications for problems in signal processing such as compressed sensing and spike train deconvolution.
In the second part, we consider the problem of factorizing a given matrix into two factors of low rank, out of which one is binary. We devise a provably correct algorithm computing such factorization whose running time is exponential only in the rank of the factorization, but linear in the dimensions of the input matrix. Our approach is extended to noisy settings and applied to an unmixing problem in DNA methylation array analysis. On the theoretical side, we relate the uniqueness of the factorization to Littlewood-Offord theory in combinatorics.

Zusammenfassung

Fortschritte in Informations- und Messtechnologie führen zu erhöhtem Vorkommen hochdimensionaler Daten. Modellierungsansätze basierend auf Sparsity oder niedrigem Rang können als Dimensionsreduktion betrachtet werden, die notwendig ist, um kompakte und interpretierbare Darstellungen solcher Daten zu erhalten.
In dieser Arbeit untersuchen wir Aspekte dieser Ansätze in Verbindung mit nichtnegativen Daten oder Nichtnegativitätsbeschränkungen.
Der erste Teil handelt vom Lernen nichtnegativer sparsamer Darstellungen, mit einem Schwerpunkt darauf, wie Nichtnegativität ausgenutzt werden kann. Wir analysieren nichtnegative kleinste Quadrate im Detail und zeigen, dass unter gewissen Bedingungen Sparsity-fördernde Regularisierung - der in den letzten Jahren paradigmatisch enpfohlene Ansatz - nicht notwendig ist. Unsere Resultate haben Auswirkungen auf Probleme in der Signalverarbeitung wie Compressed Sensing und die Entfaltung von Pulsfolgen.
Im zweiten Teil betrachten wir das Problem, eine Matrix in zwei Faktoren mit niedrigem Rang, von denen einer binär ist, zu zerlegen. Wir entwickeln dafür einen Algorithmus, dessen Laufzeit nur exponentiell in dem Rang der Faktorisierung, aber linear in den Dimensionen der gegebenen Matrix ist. Wir erweitern unseren Ansatz für verrauschte Szenarien und wenden ihn zur Analyse von DNA-Methylierungsdaten an. Auf theoretischer Ebene setzen wir die Eindeutigkeit der Faktorisierung in Beziehung zur Littlewood-Offord-Theorie aus der Kombinatorik.

Publication record

Many parts of this thesis have been published before. We here provide an overview on the underlying publications.

§1.3, 1.4 are to great parts based on:

M. Slawski and M. Hein.
Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization.
The Electronic Journal of Statistics, 7:3004–3056, 2013.

M. Slawski and M. Hein.
Sparse recovery by thresholded non-negative least squares.
Advances in Neural Information Processing Systems, 24:1926–1934. 2011.

M. Slawski and M. Hein.
Robust sparse recovery with non-negativity constraints.
4th Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS), p.30, 2011.

§1.5 is based on:

M. Slawski, R. Hussong, A. Tholey, T. Jakoby, B. Gregorius, A. Hildebrandt, and M. Hein.
Peak pattern deconvolution for protein mass spectrometry by non-negative least squares/least absolute deviation template matching.
BMC Bioinformatics, 13:291, 2012.

M. Slawski and M. Hein.
Practical Applications of Sparse Modeling, edited by I. Rish, G. Cecchi, A. Lozano and A. Niculecu-Mizil, chapter 'Sparse Recovery for Protein Mass Spectrometry Data'.
MIT press. In press.

§2 is based on:

M. Slawski, M. Hein, and P. Lutsik.
Matrix Factorization with Binary Components.
Advances in Neural Information Processing Systems. 26:3210–3218, 2013.

During and before the work on this thesis, the author published several additional papers whose results are not presented here.

M. Slawski and M. Hein.
Positive definite M-matrices and structure learning in attractive Gaussian Markov random fields.
Linear Algebra and its applications. In press.

M. Slawski
The structured elastic net for quantile regression and support vector classification.
Statistics and Computing, 22, 153-168, 2012.

M. Slawski, W. zu Castell, and G. Tutz
Feature Selection Guided by Structural Information.
Annals of Applied Statistics, 4, 1056-1080, 2010.

A.-L. Boulesteix and M. Slawski
Stability and Aggregation of ranked gene lists.
Briefings in Bioinformatics, 10, 556-568, 2009.

M. Slawski, M. Daumer, and A.-L. Boulesteix
CMA - a comprehensive Bioconductor package for supervised classification with high-dimensional data
BMC Bioinformatics, 9:439, 2008

Prologue

We have reached an era in which it is easy to collect, store, access and disseminate large data sets. While the information contained therein may carry a lot of potential for science, engineering and business, the analysis of such data poses new challenges such as massive sample size or high dimensionality. The latter refers to the availability of many attributes ('variables') per datum which can be critical, among others, for the following reasons.

Lack of interpretability: the results of data analysis tend to be difficult to interpret unless they involve a suitable condensed representation of the given data.

Reduced statistical performance: it is well documented in existing literature that several traditional data analysis techniques routinely used in a low-dimensional setting exhibit poor statistical performance when applied to data sets for which the ratio of the sample size and the number of variables is small. This situation bears the danger of overfitting, or more generally, noise accumulation ([60], §3.2).

High computing times: it is clear that performing standard tasks such as regression, classification or clustering consumes more time the more variables are taken into account. In some applications their number can be in the order of millions so that e.g. prediction of future observations may become impractically slow.

These issues indicate that it is worthwhile to consider some form of (linear) dimension reduction as an integral part of data analysis. In fact, high-dimensional data sets usually possess low-dimensional structure to be exploited. Sparsity and low-rank structure are among the best studied examples over the past few years. Here, sparsity refers to the situation where a given task can be tackled by identifying a comparatively small subset of relevant variables, while low-rank structure refers to a data matrix which is (effectively) of low rank, or in geometrical terms, to data points residing approximately in a low-dimensional linear subspace. In this thesis, we investigate aspects of both concepts in the presence of non-negative data or non-negativity constraints. A major portion is devoted to the analysis of non-negative least squares, an optimization problem one encounters when fitting linear models with non-negative parameters of the form

$$y \approx X_1\beta_1^* + \ldots + X_p\beta_p^*, \qquad \beta_j^* \geq 0, \;\; j = 1, \ldots, p, \tag{0.1}$$

where $y = (y_i)_{i=1}^n$ represents a set of observations linked to given *explanatory* or *predictor* variables X_j, $j = 1, \ldots, p$. The standard method for inferring the parameters $\{\beta_j^*\}_{j=1}^p$ is by minimizing the least squares criterion

$$\min_{\{\beta_j\}_{j=1}^p} \|y - X_1\beta_1 - \ldots - X_p\beta_p\|_2^2.$$

In case the parameters are known to be non-negative, it is recommended to impose corresponding constraints. For example, model (0.1) is suited to situations where the observations arise from an addition of certain components whose abundances are quantified in terms of the $\{\beta_j^*\}_{j=1}^p$. The setting in which *both* the parameters and the explanatory variables are non-negative is studied in greater depth in this thesis. Our specific interest is motivated by the fact that a good deal of contemporary data such as binary (0/1) data, counts or intensities (e.g. greyscale images) have a non-negative range. A distinctive feature of the two-fold non-negativity is that terms in the sum in (0.1) can no longer cancel out. This property bears the potential to curb overfitting. It also turns out to have a remarkably positive effect under sparse scenarios in which most of the $\{\beta_j^*\}_{j=1}^p$ are zero or of negligible magnitude. More specifically, the explicit promotion of sparsity by means of a data-independent regularization term as made popular by methods like the lasso [36, 151], may no longer be required. This opens up a conceptually much simpler approach to sparse recovery, i.e. identification of the set of relevant predictor variables $\{j : \beta_j^* \neq 0\}$.
The power of non-negativity was recognized earlier, but solid theoretical evidence for empirical observations has been scarce − a gap that we try to bridge in this thesis. Our findings also shed some light on non-negative matrix factorization (NMF), a popular method of linear dimension reduction for non-negative data. Given a set of non-negative data points $d_j \in \mathbb{R}_+^m$, $j = 1, \ldots, n$, NMF aims at finding a set of points $t_k \in \mathbb{R}_+^m$, $k = 1, \ldots, r$, with $r < \min\{m, n\}$ chosen according to the desired dimension reduction such that

$$d_j \approx t_1\alpha_{1j} + \ldots + t_r\alpha_{rj} \tag{0.2}$$

for non-negative coefficients α_{kj}, $j = 1, \ldots, n$, $k = 1, \ldots, r$. Note that for each j, (0.2) constitutes a linear model of the form (0.1), with the important difference that there are no fixed predictor variables; the $\{t_k\}_{k=1}^r$ and the coefficients $\{\alpha_{kj}\}$ need to be inferred simultaneously. This can be recast as the optimization problem

$$\min_{T\in\mathbb{R}_+^{m\times r},\, A\in\mathbb{R}_+^{r\times n}} \|D - TA\|_F^2 = \min_{T\in\mathbb{R}_+^{m\times r},\, A\in\mathbb{R}_+^{r\times n}} \sum_{i=1}^m \sum_{j=1}^n (D_{ij} - (TA)_{ij})^2, \tag{0.3}$$

where the columns of the matrices D and T contain the $\{d_j\}_{j=1}^n$ respectively the $\{t_k\}_{k=1}^r$ and $A = (\alpha_{kj})$. In their seminal paper [92], Lee and Seung argue that NMF has tendency to yield a parts-based decomposition of the data matrix D in which the 'parts' $\{t_k\}_{k=1}^r$ and the associated coefficients are sparse. This phenomenon can be better understood in light of our analysis of non-negative least squares under two-fold non-negativity. In fact, problem (0.3) collapses into n independent non-negative least squares problems of the form

$$\min_{\{\alpha_{kj}\geq 0\}_{k=1}^r} \|d_j - t_1\alpha_{1j} - \ldots - t_r\alpha_{rj}\|_2^2, \quad j = 1, \ldots, n,$$

once the matrix T is known (and vice versa if A is known). This observation also underlies the common alternating updates approach to optimize the NMF criterion (0.3). However, such scheme cannot be shown to yield a global minimizer. Indeed, solving (0.3) globally optimally yields a computational challenge in general. Specifically, even in the exact case in which it holds that $D = TA$ with T and A non-negative, finding

such factorization has been shown to be NP-hard in [160]. The influential paper by Arora et al. [3] has given fresh impetus to the field. In that paper, the authors show that the exact NMF problem can be solved by linear programming under a certain condition fulfilled in several important applications of NMF.
In the second chapter of this thesis, we discuss the computation of a special case of NMF in which one of the two factors is required to be binary. We show that in a wide range of cases, solving NMF problems with one binary factor remains computationally tractable as long as the inner dimension r of the factorization remains small. This may come as a surprise since − at least at first glance − the additional combinatorial constraints seem to add another layer complexity to an already challenging problem.
In summary, we hope to convey the idea that non-negativity is a common yet powerful constraint that can be of enormous use in data analysis, and that offers various interesting directions of research.

Chapter 1

Sparse recovery with non-negativity constraints

The problem of learning sparse representations from few samples has received enormous attention in the past ten to fifteen years in a variety of disciplines in mathematics, engineering, and computer science. Interest in this problem has evolved from the practical need to find compact representations of high-dimensional objects (e.g. images or videos), in order to enable efficient compression and sampling [18, 144], as well as from the search for statistically sound methods dealing with high-dimensional feature spaces [19, 82]. In this work, we study sparse, *non-negative* representations, with a specific focus on how the additional sign constraint can be harnessed in terms of statistical theory and practical data analysis.

Chapter outline. We start by providing an overview on high-dimensional linear models and sparse estimation, the theme of this chapter. It consists of a theoretical and a practical part. The theoretical part follows a division into a noiseless and a noisy setup. For the former, we consider the problem of recovering a sparse non-negative vector from underdetermined linear systems of equations and discuss the underlying geometry. Our treatment of the noisy setup focuses on a detailed analysis of non-negative least squares within a modern framework of high-dimensional statistical inference. The practical part is devoted to an in-depth case study in proteomics, which is used to highlight the usefulness of non-negative least squares in applications. A list of contributions of this chapter can be found in §1.1.5.

Notation table for this chapter.

$:=$	equality by definition; we use plain '=' if context is clear
$I(\cdot)$	indicator function
v_I	subvector of $v \in \mathbb{R}^m$ corresponding to $I \subseteq \{1, \ldots, m\}$
$\|v\|_q$	ℓ_q-(quasi)-norm of $v \in \mathbb{R}^m$, i.e. $\|v\|_q = (\sum_{i=1}^m v_i^q)^{1/q}$ for $q \in (0, \infty)$ and $\|v\|_\infty = \max_{1 \le i \le m} \|v_i\|$
$\|v\|_0$	ℓ_0-'norm' of v, i.e. $\|v\|_0 = \sum_{i=1}^m I(v_i \neq 0)$
S^c	complement of S (if ground set is clear from the context)
$\|S\|$	cardinality of a set S
$B_0(S; p)$	vectors in $\mathbb{R}^p$ with support included in $S \subseteq \{1, \ldots, p\}$, i.e. $B_0(S; p) = \{v \in \mathbb{R}^p : v_j = 0 \ \forall j \notin S\}$
$B_0(s; p)$	vectors in $\mathbb{R}^p$ with at most s non-zero entries, i.e. $B_0(s; p) = \{v \in \mathbb{R}^p : \|v\|_0 \le s\}$
$\mathbb{R}_+$	non-negative real line, i.e. $\mathbb{R}_+ = \{x \in \mathbb{R} : x \ge 0\}$
$B_0^+(S; p)$	$B_0(S; p) \cap \mathbb{R}_+^p$
$B_0^+(s; p)$	$B_0(s; p) \cap \mathbb{R}_+^p$
$\mathcal{J}(k)$, $0 \le k \le p$	sets of all subsets of $\{1, \ldots, p\}$ of cardinality k, i.e. $\mathcal{J}(k) = \{J \subseteq \{1, \ldots, p\} : \|J\| = k\}$
$\langle v, w \rangle$	usual inner product of $v, w \in \mathbb{R}^m$, i.e. $\langle v, w \rangle = \sum_{i=1}^m v_i w_i$
$v \succeq w$	$v_i \ge w_i$, $i = 1, \ldots, m$. Analogous for $\preceq$, $\succ$, $\prec$.
$x \vee y$, $x \wedge y$	$\max\{x, y\}$, $\min\{x, y\}$
$\lfloor x \rfloor$	largest integer less than or equal to x
T^{m-1}	probability simplex in $\mathbb{R}_+^m$, i.e. $T^m = \{x \in \mathbb{R}_+^m : \sum_{i=1}^m x_i = 1\}$
int A	interior of some set $A \subseteq \mathbb{R}^m$ (w.r.t. the usual topology)
bd A	boundary of some set $A \subseteq \mathbb{R}^m$ (w.r.t. the usual topology)
conv A	convex hull of some set $A \subseteq \mathbb{R}^m$

M_{IJ}	submatrix of a real $m \times n$ matrix M corresponding to rows in $I \subseteq \{1, \ldots, m\}$ and columns in $J \subseteq \{1, \ldots, n\}$
M_J	submatrix of M corresponding to columns in J; $M_\emptyset := 0$
M_j	j-th column of M
M^j	(transpose of the) j-th row of M
$[M,\, M']$	column-wise concatenation of matrices M, M'
$[M;\, M']$	row-wise concatenation of matrices M, M'
$\mathcal{N}(M)$	nullspace of a real matrix M, i.e. $\mathcal{N}(M) = \{x \in \mathbb{R}^n : Mx = 0\}$
$\mathrm{tr}(M)$	trace of a square matrix M, i.e. $\mathrm{tr}(M) = \sum_{i=1}^m M_{ii}$
$\mathcal{C}_M$	conic hull of the columns of M, i.e. $\mathcal{C}_M = \{y \in \mathbb{R}^m :\ y = M\lambda,\ \lambda \in \mathbb{R}^n_+\}$
$\mathcal{P}_M$	convex hull of the columns of M, i.e. $\mathcal{P}_M = \{y \in \mathbb{R}^m :\ y = M\lambda,\ \lambda \in \mathbb{R}^n_+, \sum_{i=1}^n \lambda_i = 1\}$
$\mathcal{P}_{0,M}$	convex hull of the columns of M and the origin, i.e. $\mathcal{P}_{0,M} = \{y \in \mathbb{R}^m :\ y = M\lambda,\ \lambda \in \mathbb{R}^n_+, \sum_{i=1}^n \lambda_i \leq 1\}$
I_m I	identity matrix of dimension m, identity matrix of unspecified dimension (context)
$\{e_i\}_{i=1}^m$	canonical basis vectors of $\mathbb{R}^m$
$\mathbf{1}$, $\mathbb{1}$	vector respectively matrix of ones
$\mathbf{E}[Z]$	expectation of some random variable Z
$\mathbf{Var}[Z]$	variance of some random variable Z
$\mathbf{P}(A)$	probability of some event A
$Z \sim N(\mu, C)$	random vector Z follows a Gaussian distribution with mean μ and covariance C
$Y \stackrel{\mathcal{D}}{=} Z$	the random variables Y and Z follow the same distribution

$f(x) = o(g(x))$ as $x \to a$	$\lim_{x \to a} \lvert f(x)/g(x) \rvert = 0$
$f(x) = O(g(x))$ as $x \to a$	$\limsup_{x \to a} \lvert f(x)/g(x) \rvert < \infty$
$f(x) = \Omega(g(x))$ as $x \to a$	$\liminf_{x \to a} \lvert f(x)/g(x) \rvert > 0$
$f(x) = \Theta(g(x))$ as $x \to a$	$0 < \liminf_{x \to a} \lvert f(x)/g(x) \rvert \leq \limsup_{x \to a} \lvert f(x)/g(x) \rvert < \infty$
$X_n = o_{\mathbf{P}}(g(n))$	Sequence of random variables $\{X_n\}$ satisfies $\forall \varepsilon > 0 \ \lim_{n \to \infty} \mathbf{P}(\lvert X_n/g(n) \rvert \geq \varepsilon) = 0.$
$X_n = O_{\mathbf{P}}(g(n))$	Sequence of random variables $\{X_n\}$ satisfies $\forall \delta > 0 \ \exists c < \infty$ such that $\mathbf{P}(\lvert X_n/g(n) \rvert \geq c) < \delta \ \ \forall n.$
c, c', c_1, C, C', C_1 etc.	positive constants (value may differ from line to line)

Abbreviations, acronyms, and terminology

i.i.d.	independent identically distributed
(l/r).h.s.	(left/right) hand side
MSE	mean squared error
NNLS	non-negative least squares
PSF	point spread function
s.t.	such that
sb. to	subject to
w.r.t.	with respect to

'Random vector X is standard Gaussian': short for $X \sim N(0, I)$

1.1. Background on sparse recovery and statistical estimation for sparse high-dimensional linear models

In this section, we give a general overview on the topic under consideration and discuss its relations to several areas of recent research.

1.1.1 Problem statement

For what follows, we consider *observations* $y \in \mathbb{R}^n$, which are modelled according to the linear model

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

where the right hand side is composed of the following quantities.

- $X \in \mathbb{R}^{n\times p}$ is referred to as *design matrix*. Depending on the context, X may be regarded as fixed (*fixed design*) or random (*random design*).
- $\beta^* \in \mathbb{R}^p_+$ is a *non-negative* vector with *support* $S = \{j : \beta^*_j > 0\}$. We write $s = |S|$ for its cardinality, to which we refer as *sparsity level*.
- ε represents an error term whose components $(\varepsilon_i)_{i=1}^n$ are typically i.i.d. random ('noise') variables. If $\varepsilon = 0$, we speak of the *noiseless case* and otherwise of the *noisy case*.

The goal is to infer the unknown parameter β^* given (X, y), i.e. one wants to find an estimator $\widehat{\theta} = \widehat{\theta}(X, y)$ that estimates accurately the *target* β^* contained in the parameter set $B_0^+(s; p) = \{\beta \in \mathbb{R}^p_+ : \|\beta\|_0 \le s\}$ with respect to one of the following criteria.

Exact recovery. In the noiseless case, one mostly aims at having $\widehat{\theta} = \beta^*$.

Estimation error. In the noisy case, exact recovery is not possible in general. A natural measure for the goodness of approximation is the error in ℓ_q-norm $\|\widehat{\theta} - \beta^*\|_q$ for some $q \in [1, \infty]$.

Prediction error. In the noisy case, one may be interested in reducing the contamination of the observations by the noise ε (*denoising*). A common measure is the mean squared (prediction) error (MSE) $\frac{1}{n}\|X\widehat{\theta} - X\beta^*\|_2^2$. The term 'prediction error' here stems from the fact that for deterministic X, this quantity reveals how well on average $X\widehat{\theta}$ predicts a new set of observations $\widetilde{y} = X\beta^* + \widetilde{\varepsilon}$ if $\widetilde{\varepsilon}$ is a zero-mean random vector independent of ε:

$$\begin{aligned}\mathbf{E}\left[\frac{1}{n}\|\widetilde{y} - X\widehat{\theta}\|_2^2 \,\middle|\, y\right] = \mathbf{E}\left[\frac{1}{n}\|X\beta^* - X\widehat{\theta} + \widetilde{\varepsilon}\|_2^2 \,\middle|\, y\right] &= \frac{1}{n}\|X\widehat{\theta} - X\beta^*\|_2^2 + \mathbf{E}\left[\frac{1}{n}\|\widetilde{\varepsilon}\|_2^2\right] \\ &= \frac{1}{n}\|X\beta^* - X\widehat{\theta}\|_2^2 + \text{const.},\end{aligned}$$

because the second term does not depend on $\widehat{\theta}$. For the second identity, we have invoked the assumption that $\mathbf{E}[\widetilde{\varepsilon}] = 0$. As discussed in subsequent sections, the MSE is a suitable criterion if it is not possible to derive a reasonable upper bound on the estimation error due to high correlations among subsets of columns of X. On the other hand, an upper bound on the ℓ_2-estimation error trivially yields a bound on the MSE:

$$\frac{1}{n}\|X\widehat{\theta} - X\beta^*\|_2^2 \leq \phi_{\max}\left(\frac{1}{n}X^\top X\right)\|\widehat{\theta} - \beta^*\|_2^2,$$

where $\phi_{\max}(M)$ denotes the largest eigenvalue of a real symmetric matrix M.

Support recovery or sign consistency. In a sparse regime in which the sparsity level of β^* is substantially smaller than the dimension, it is desirable to have a sparse estimator $\widehat{\theta}$ whose support $\{j : \widehat{\theta}_j \neq 0\}$ agrees with that of β^* (i.e. $\widehat{\theta}$ recovers the support of β^*), because it means that one has achieved a correct reduction in model complexity. This aspect roots in the problem of *variable* or *feature selection* in linear regression [113]. Sign-consistency is a slightly more stringent notion than support recovery, which requires that $\operatorname{sign}(\widehat{\theta}_j) = \operatorname{sign}(\beta_j^*)$, $j = 1, \ldots, p$. Clearly, if the estimator $\widehat{\theta}$ takes values in $\mathbb{R}_+^p$, the two notions coincide.

1.1.2 The high-dimensional, sparse setting

In the present work, we focus on the 'high-dimensional, (but) sparse regime' of modern statistical theory ([19], §1), which is outlined in the sequel. Classical statistical estimation theory studies the behaviour of an estimator for a fixed parameter set while the sample size n tends to infinity. This framework has lost in importance because it does not cover several other regimes relevant to modern data analysis. As to our problem introduced above, consider the least squares estimator

$$\widehat{\beta}^{\mathrm{ols}} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n}\|y - X\beta\|_2^2. \tag{1.2}$$

Suppose that X is deterministic and, for simplicity, that ε has i.i.d. zero-mean Gaussian entries. Let Π_X denote the projection on the column space of X of dimension $d = \operatorname{tr}(\Pi_X)$. From $X\widehat{\beta}^{\mathrm{ols}} = \Pi_X y = X\beta^* + \Pi_X\varepsilon$, we obtain the following for the prediction error of $\widehat{\beta}^{\mathrm{ols}}$

$$\frac{1}{n}\|X\widehat{\beta}^{\mathrm{ols}} - X\beta^*\|_2^2 = \frac{1}{n}\|\Pi_X\varepsilon\|_2^2 = O_{\mathbf{P}}(d/n) = O_{\mathbf{P}}(p/n). \tag{1.3}$$

We may draw the conclusion that the prediction error of $\widehat{\beta}^{\mathrm{ols}}$ vanishes asymptotically as $n \to \infty$ while p stays fixed. On the other hand, such asymptotic consideration is not meaningful from a practical point of view if the given data (X, y) do not well reflect this scenario because p is actually similarly large as n or even larger. The fact that the prevalence of such data has dramatically increased in recent years (cf. the following subsection), together with considerable advance in theory, has led to a novel framework whose innovations are summarized below.

Increasing sequence of parameter sets. The parameter set is allowed to increase with n. For the problem under consideration with parameter set $B_0^+(s;p)$ this means that the problem dimension $p = p_n$, the sparsity level $s = s_n$ and thus $\beta^* = \beta_n^*$ may depend on n. Accordingly, asymptotics are understood with respect to a a triangular array of observations $\{(X^{(n)}, y^{(n)}), X^{(n)} \in \mathbb{R}^{n \times p_n}\}, \quad n = 1, 2, \dots$ For notational simplicity, dependence on n is usually suppressed. Therefore, we here stress that any quantity depending on n via p, s or β^* can, in general, no longer be regarded as a constant.

Non-asymptotic analysis. A non-asymptotic analysis, in which results are stated for finite sample sizes, is preferred, even though one typically thinks of n being large when interpreting these results.

Leveraging sparsity. In order to establish reasonable performance guarantees even if $p > n$, one needs to assume that the problem is sufficiently sparse, i.e. that s is small relative to n. In this case, the set $B_0(s;p) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s\}$, which is the union of all subspaces spanned by selections of s canonical basis vectors, is effectively a low-dimensional object. Moreover, one needs to work with estimators that are able take advantage of such structure.
We remark that the assumption of *exact sparsity*, i.e. sparsity in an ℓ_0-sense, can be relaxed to various forms of *approximate sparsity*, in which β^* is not required to have few non-zero entries but instead few entries of significant magnitude. Common models of approximate sparsity can be found in [124], §2.1. We restrict ourselves to exact sparsity, apart from a single theorem (Theorem 1.28).

Scalings of p and s: regimes of interest. We now discuss specific scalings of p and s with n that have been frequently considered in the literature and which are of interest in this work. When speaking of a 'high-dimensional setting', we suppose that we have at least $p = \Theta(n)$ up to $p = o(\exp(n^\xi))$ for some $0 < \xi < 1$. Note that for $p = o(n)$, even ordinary least squares tends to perform well (cf. (1.3)), hence this regime is not of specific interest. In the noiseless case, we restrict our attention to the situation $n < p$. Regarding the scaling of s, the regime $p = \Theta(n) = \Theta(s)$, $s < n < p$, which is typically referred to as *regime of linear sparsity* [163] or *proportional growth setting* [51], is of particular interest in the noiseless case. In the noisy case, it is standard to assume sublinear sparsity in the sense that $s = o(n/\log p)$.

1.1.3 Practical relevance of the high-dimensional, sparse setting

The high-dimensional, but sparse setting is not only a construct for theoretical analysis. In fact, such setting is ubiquitous in contemporary applications, a selection of which is outlined below.

Learning from many features. The linear model (1.1) arises in linear regression, where the goal is to model an outcome variable as a linear combination of *(input) variables*, *predictors* or *features*. The high-dimensional setting has gained in importance in this context, among others, for the following reasons.

- Since the advent of the digital age and the associated advances in information technology, it is cheap to collect and store many attributes per observation. The hope is that the more information, i.e. the more attributes, are taken into account, the more accurately an outcome variable of interest can be predicted.
- It is common to augment the set of given features by additionally considering nonlinear transformations thereof, e.g. powers, logarithm, or products of several features (see pp.139-141 in [74] for more examples). The thus 'enriched' feature set is supposed to yield more flexibility in modelling and in turn also improved performance in prediction. At the same time, the number of features p may grow quickly depending on the transformations considered. For example, if $\underline{p}$ denotes the original number of features and one considers all products involving up to d features, we end up with $p > (\underline{p}/d)^d$ features.
- In data from high-throughput biological experiments, for example gene expression microarrays, it is common to have few observations but many features ($n \ll p$ set-up). We refer to [10] for an overview.

Sparsity, on the other hand, is a reasonable assumption as long as only a small fraction of all predictors considered have a significant effect on the outcome variable. Moreover, sparse models are desired from the points of view of interpretation and computation (e.g. in order to reduce time and storage requirements for prediction).

Sparse approximation with overcomplete systems. This topic has received much attention in mathematical signal processing (see e.g. [18], §4.2 and §4.3) and concerns the sparse representation of a given signal $y \in \mathbb{R}^n$ in a union of bases of $\mathbb{R}^n$. This model is suitable whenever the signal arises from a superposition of heterogeneous components.

Inverse problems. The matrix X may also represent operations that lead to a degraded version y of some underlying signal β^*. For example y may be a blurred version of some image β^*. Within this thesis, specific attention is paid to sparse spike train deconvolution, where the locations of the non-zero entries in β^* indicate the positions of spikes and X represents convolution of the spike train with a *point-spread function (PSF)* and possible down-sampling (cf. §1.5). Typically, problems of this kind fall into the regime $p = \Theta(n)$.

Compressed sensing. Compressed sensing (CS) is a modern sampling paradigm in signal processing pioneered in [31, 33, 45] from which an entire new field of research has emerged, see [56, 124] for an overview. The goal of CS is to recover a signal $\beta^* \in \Theta \subseteq \mathbb{R}^p$ from few samples (one often speaks of *measurements* in this context), where the process of sampling or at least aspects thereof can be designed by the user. CS consists of two basic steps termed *sampling* and *decoding*. In the sampling step, one obtains measurements $y_i = g_i(\beta^*; \{y_u\}_{u<i})$, $i = 1, \ldots, n$, $n \ll p$, for functions $g_i : \Theta \to \mathbb{R}$ that may depend on all preceding measurements $\{y_u\}_{u<i}$, $i = 1, \ldots, n$. The decoding step consists of a mapping $\Delta : \mathbb{R}^n \to \Theta$ whose goal is to recover β^* from the measurements y. While in general the measurement process may be both non-linear (i.e. the functions $\{g_i\}_{i=1}^n$ may be non-linear in β^*) and adaptive (the functions $\{g_i\}_{i=1}^n$

may be chosen depending on earlier measurements), it is the case of linear and non-adaptive measurements with $y_i = \langle X^i, \beta^* \rangle$, $X^i \in \mathbb{R}^p$, $i = 1, \ldots, n$ that has received most attention in the literature. Note that this case is subsumed by our linear model (1.1) with the $\{X^i\}_{i=1}^n$ stacked into the rows of the matrix X (the error term ε can be used to model noise in the measurement process), and the decoding step of CS becomes a special case of the problem under consideration. A crucial difference from the setups in regression or deconvolution discussed above is that in CS, the matrix X is regarded as an object that may be chosen freely from the set of $n \times p$ real matrices. In particular, various random constructions of X were considered already at early stages of CS.

1.1.4 Estimation procedures for sparse high-dimensional linear models

Before discussing the peculiarities of the sparse, non-negative case with $\beta^* \in B_0^+(s;p)$, which is in the center of this thesis, we first provide a short survey on the sparse case with $\beta^* \in B_0(s;p)$. If s is known, a straightforward approach directly incorporating the given prior knowledge about the target is ℓ_0-constrained least squares estimation which yields the estimator

$$\widehat{\beta}^{\ell_0,s} \in \operatorname*{argmin}_{\beta \in B_0(s;p)} \frac{1}{n} \|y - X\beta\|_2^2. \tag{1.4}$$

Under the condition $\phi_{\min}(2s) > 0$, where for $k \in \{1, \ldots, p\}$

$$\phi_{\min}(k) = \min_{\delta \in B_0(k;p) \setminus \{0\}} \frac{\|X\delta\|_2^2}{n\|\delta\|_2^2}, \tag{1.5}$$

the following statement establishes several performance guarantees for $\widehat{\beta}^{\ell_0,s}$. Our proof closely follows the proof of Theorem 2 in [123].

Proposition 1.1. *Consider the linear model (1.1) with $\beta^* \in B_0(s;p)$ and support $S = \{j : \beta_j^* \neq 0\}$. Suppose that $\phi_{\min}(2s) > 0$ and denote $A = \max_{1 \le j \le p} |X_j^\top \varepsilon|/n$. We then have*

$$\|\widehat{\beta}^{\ell_0,s} - \beta^*\|_q^q \le \frac{2^{q+1} A^q s}{\{\phi_{\min}(2s)\}^q}, \ q \in [1,2], \quad \text{and} \quad \frac{1}{n}\|X\widehat{\beta}^{\ell_0,s} - X\beta^*\|_2^2 \le \frac{8 A^2 s}{\phi_{\min}(2s)}. \tag{1.6}$$

Furthermore, if $\min_{j \in S} |\beta_j^*| > \frac{(2A\sqrt{2s})}{\{\phi_{\min}(2s)\}}$, it holds that $sign(\widehat{\beta}_j^{\ell_0,s}) = sign(\beta_j^*)$, $j = 1, \ldots, p$.

Proof. From the definition of $\widehat{\beta}^{\ell_0,s}$, we get

$$\frac{1}{n}\|y - X\widehat{\beta}^{\ell_0,s}\|_2^2 \le \frac{1}{n}\|y - X\beta^*\|_2^2.$$

Let $\delta = \widehat{\beta}^{\ell_0,s} - \beta^*$. Using that $y = X\beta^* + \varepsilon$, it follows that

$$\frac{1}{n}\|X\delta\|_2^2 \le \frac{2}{n}\left|\left\langle \delta, X^\top \varepsilon \right\rangle\right| \le 2A\|\delta\|_1, \tag{1.7}$$

where the second inequality results from Hölder's inequality and the definition of A. Noting that $\delta \in B_0(2s;p)$, we obtain according to the definition of $\phi_{\min}(2s)$

$$\phi_{\min}(2s)\|\delta\|_2^2 \leq \frac{1}{n}\|X\delta\|_2^2 \leq 2A\|\delta\|_1$$

The fact that $\delta \in B_0(2s;p)$ also implies that $\|\delta\|_2^2 \geq \|\delta\|_1^2/2s$ and thus

$$\|\widehat{\beta}^{\ell_0,s} - \beta^*\|_1 = \|\delta\|_1 \leq \frac{4A\,s}{\phi_{\min}(2s)}, \quad \text{and} \quad \|\widehat{\beta}^{\ell_0,s} - \beta^*\|_2^2 = \|\delta\|_2^2 \leq \frac{8A^2\,s}{\{\phi_{\min}(2s)\}^2} \tag{1.8}$$

Substituting the bound on $\|\delta\|_1$ back into (1.7), we obtain the second part of (1.6). The general ℓ_q-bound results from the inequality $\|\delta\|_q^q \leq \|\delta\|_1^{2q-1}\|\delta\|_2^{2(q-1)}$, which holds for all $q \in [1,2]$. As to the second part of the statement, we have from (1.8)

$$\frac{2A\sqrt{2s}}{\phi_{\min}(2s)} \geq \|\widehat{\beta}^{\ell_0,s} - \beta^*\|_2 \geq \|\widehat{\beta}^{\ell_0,s} - \beta^*\|_\infty \geq \max_{j\in S}|\widehat{\beta}_j^{\ell_0,s} - \beta_j^*|$$

If there were a $j \in S$ such that $\operatorname{sign}(\widehat{\beta}_j^{\ell_0,s}) \neq \operatorname{sign}(\beta_j^*)$, the lower bound on $\min_{j\in S}|\beta_j^*|$ would lead to a contradiction. Consequently, we must have $\operatorname{sign}(\widehat{\beta}_j^{\ell_0,s}) = \operatorname{sign}(\beta_j^*)$, $j \in S$ and in turn $\|\widehat{\beta}^{\ell_0,s}\|_0 = s$ and hence also $\widehat{\beta}_j^{\ell_0,s} = 0$ for all $j \in S^c$. □

Proposition 1.1 reveals that from a mere statistical point of view, ℓ_0-constrained least squares allows one to cope with the high-dimensional, sparse setting. All bounds depend on p only via $\phi_{\min}(2s)$ (it turns out not to be restrictive to assume the scaling $\phi_{\min}(2s) = \Omega(1)$) and the term A that represents the influence of the error term. Specializing to $\varepsilon = 0$, Proposition 1.1 asserts exact recovery, i.e. $\widehat{\beta}^{\ell_0,s} = \beta^*$. If ε has i.i.d. zero-mean Gaussian entries and $\max_{1\leq j\leq p}\|X_j\| = O(\sqrt{n})$, one can show that $A = O_{\mathbf{P}}(\log(p)/n)$ (cf. the proof of Theorem 1.21 below). The bounds (1.6) then yield

$$\|\widehat{\beta}^{\ell_0,s} - \beta^*\|_2^2 = O_{\mathbf{P}}(s\log(p)/n), \quad \text{and} \quad \frac{1}{n}\|X\widehat{\beta}^{\ell_0,s} - X\beta^*\|_2^2 = O_{\mathbf{P}}(s\log(p)/n). \tag{1.9}$$

The bound on the prediction error constitutes a drastic improvement over the corresponding bound for least squares estimation (1.3). In contrast to the latter, the bound for $\widehat{\beta}^{\ell_0,s}$ reflects the sparsity of the problem with linear dependence on s in place of p, which now enters only logarithmically. Apart from the extra log factor, the bounds (1.9) match the performance of an estimator one would use if one had access to an oracle revealing the support of β^*:

$$\widehat{\beta}^{\text{oracle}} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p:\beta_{S^c}=0} \frac{1}{n}\|y - X\beta\|_2^2. \tag{1.10}$$

The second part of Proposition 1.1 implies that if all non-zero entries of β^* are sufficiently large, then $\widehat{\beta}^{\ell_0,s} = \widehat{\beta}^{\text{oracle}}$. However, in practice it is basically as unrealistic to use $\widehat{\beta}^{\ell_0,s}$ as is the presence of an oracle. Let us detail on this.

- ℓ_0-constrained least squares estimation is a non-adaptive estimation procedure in the sense that the sparsity level of the problem needs to be known in advance in order to achieve performance bounds of the correct order. In contrast, an adaptive estimation procedure achieves optimal performance simultaneously over a broad range of sparsity levels. In [23] it is shown that under assumption of i.i.d. zero-mean Gaussian errors ε, adaptivity is achieved when using ℓ_0-*penalized* or *regularized least squares estimation*

$$\widehat{\beta}^{\ell_0,\lambda} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0, \;\; \lambda > 0, \tag{1.11}$$

  with proper choice of the regularization parameter λ. Unfortunately, such choice depends on the variance of the error terms, which is typically not known in practice.

- Even in case s or a suitable value of λ were known, it would still not be practical to work with (1.4) respectively (1.11) for computational reasons. Computing $\widehat{\beta}^{\ell_0,s}$ in the most obvious way involves checking all $\sum_{k=0}^{s}\binom{p}{k}$ subsets of $\{1,\dots,p\}$ of cardinality less than or equal to s, which is not feasible in practice unless s is tiny (say $s \leq 3$) or p is small (state-of-the-art branch-and-bound methods [76] may handle cases with p up to 50). From the point of view of computational complexity, several hardness results have been established [2, 114]. There do exist algorithms that can be shown to deliver $\widehat{\beta}^{\ell_0,s}$ under certain conditions on the data (X, y) (see the subsequent paragraph for examples). However, verifying these conditions is in turn NP-hard or conjectured to be NP-hard [152].

Overall, the discussion of ℓ_0-constrained estimation has touched upon crucial criteria based on which different estimation procedures should be compared.

- What performance guarantees regarding prediction, estimation and support recovery can be established ?
- What conditions on X and $\min_{j\in S}|\beta_j^*|$ are required to achieve these guarantees, and are these conditions likely to be fulfilled for a given problem ?
- What is the computational effort needed to compute the estimator ?
- What is the degree of adaptivity of the procedure, i.e. which tuning parameters need to be specified and can these tuning parameters be chosen in a data-driven manner without explicit knowledge of problem-specific quantities ?

In the sequel, we discuss a selection of estimation procedures proposed in the literature in light of the considerations above.

Practical approaches to (approximate) ℓ_0-constrained or regularized estimation. In the preceding discussion, we have thought of (1.4) and (1.11) as obtaining the globally optimal solution to a combinatorial optimization problem. A different approach is to treat (1.4) as an instance of nonlinear programming and then apply an algorithm

from this field that allows one to circumvent the combinatorial nature of the problem. The use of gradient projection is most prominent in this context [14], as it can be exploited that it is trivial to compute the Euclidean projection on $B_0(s;p)$. This yields a practical scheme, for which performance guarantees can be established if X satisfies certain forms of the *restricted isometry property* originally introduced in [30, 31]. Roughly speaking, this condition requires that $\frac{1}{n}X^\top X$ nearly acts as an isometry on $B_0(2s;p)$, which is much more restrictive than the condition $\phi_{\min}(2s) > 0$, cf. (1.5). An alternative approach [150] applicable to both (1.4) and (1.11) is a reformulation within a certain class of optimization problems known as DC programs [42]. No performance guarantees appear to have been established for the approach in [150] so far.
Another line of research concerned with (1.11) considers families of functions that are smooth (apart from single points) and that can approximate the function $x \mapsto I(x \neq 0)$ arbitrarily well. A classical example is the family of ℓ_q-quasinorms with $x \mapsto |x|^q$, $q \in (0,1)$, see [64] and [174] for more examples. The resulting optimization problems are non-convex, which considerably complicates theoretical analysis due to the presence of multiple local optima, see [164, 174].

Convex relaxation: ℓ_1-regularization. The probably most popular approach to sparse high-dimensional regression is *ℓ_1-regularized least squares estimation*

$$\widehat{\beta}^{\ell_1,\lambda} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \quad \lambda > 0, \tag{1.12}$$

which results from (1.11) by replacing the ℓ_0-'norm' by its convex envelope[1] on $[-1,1]^p$. This motivates the use of term convex relaxation here. Convexity entails that a globally optimal solution of (1.12) can be found using one out of a whole battery of efficient algorithms [137]. ℓ_1-regularized least squares estimation has a long history in statistics [151] under the acronym 'lasso' (which will also be used here) as well as in signal processing [36]. The 'Dantzig selector' [32] is a highly similar approach. In the noiseless case, both these approaches amount to *ℓ_1-minimization*

$$\widehat{\beta}^{\ell_1} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|\beta\|_1 \quad \text{sb. to} \quad X\beta = y, \tag{1.13}$$

which coincides with (1.12) in the limit $\lambda \to 0$ provided (1.13) is feasible. By now, there is a substantial body of work [30, 31, 39, 43, 44, 46, 47, 51, 61, 132, 178] on the question of *ℓ_1/ℓ_0-equivalence* in the noiseless case, where ℓ_1-minimization is related to *ℓ_0-minimization*

$$\widehat{\beta}^{\ell_0} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \|\beta\|_0 \quad \text{sb. to} \quad X\beta = y. \tag{1.14}$$

One speaks of ℓ_1/ℓ_0-equivalence if the solutions of both (1.13) and (1.14) are unique and agree, in which case ℓ_1-minimization achieves exact recovery. One of the major findings is that for certain classes of random matrices ℓ_1/ℓ_0-equivalence holds for a wide range of scalings of the triple (n,p,s) (cf. end of §1.1.2), which underlies a remarkable high-dimensional geometric phenomenon [43, 46, 51] to be discussed in more detail for the non-negative case in §1.3.3 below.

[1] The convex envelope of a function is its tightest convex underapproximate. More precisely, it is its biconjugate [129], §12.

In the noisy case, the lasso (1.12) and ℓ_0-regularization (1.11) can no longer be exactly equivalent, but they tend to have comparable performance with regard to estimation in ℓ_q-norm, $q \in [1,2]$, and prediction error (cf. Proposition 1.1), under the conditions that, roughly speaking, the design X would satisfy ℓ_1/ℓ_0-equivalence in the noiseless case, and the regularization parameter λ is properly specified [11, 58, 110, 122, 123, 156, 157, 173, 182]. On the other hand, the situation is noticeably different from the noiseless case in the sense that the lasso does not achieve sign consistency/support recovery irrespectively of how λ is chosen, unless X satisfies a specific condition, which is rather restrictive [95, 109, 163, 180, 184]. This failure occurs irrespective of how large $\min_{j\in S}|\beta^*_j|$ is and can be traced back to the bias of the ℓ_1-regularizer that shrinks all components of $\widehat{\beta}^{\ell_1,\lambda}$ towards zero, including those corresponding to the support of β^*, where such shrinkage is actually not desired. The lasso tries to compensate for that shrinkage by including extra predictors corresponding to S^c so that $\widehat{\beta}^{\ell_1,\lambda}_{S^c}$ tends to have some entries of small, yet non-zero absolute magnitude [180]. Eventually, this issue can be seen as the price one has to pay for resorting to a relaxation, as ℓ_0-regularization is not affected by this problem. Since support recovery is of central importance in the context of variable selection, this shortcoming of the lasso has triggered much follow-up work including various suggestions on how to restore support recovery of the lasso such as the adaptive lasso [184] and the thresholded lasso [110, 181], and can still be regarded as an area of active research. A drawback of both the adaptive lasso and the thresholded lasso is that additional tuning parameters are introduced to the problem. The thresholded lasso is a two-stage procedure, in which $\widehat{\beta}^{\ell_1,\lambda}$ is obtained with a suitable choice of λ before all components of absolute magnitude below a suitably chosen threshold are set to zero. In fact, proper specification of λ is already a non-trivial task. In the case of i.i.d. zero-mean Gaussian errors, theoretical results indicate that λ should be proportional to the standard deviation of the errors, which however, is usually unknown and there is no straightforward way of estimating it [78]. In order to avoid this issue, two modifications of the lasso, the square-root lasso [7] and the scaled lasso [146] have been proposed, which achieve a similar performance as the lasso, while the correct choice of the regularization parameter does no longer depend on the standard deviation of the errors. On the other hand, these modifications, which concern the least squares term in (1.12), lead to more complicated (though still convex) optimization problems. Apart from that, the theory in [7] and [146] still involves a number of assumptions, so that both methods cannot be considered as entirely tuning-free in general. Alternatively, data-driven tuning of λ based on cross-validation (e.g. [74], §7.10) is computationally expensive and may be error-prone if λ is chosen from some grid specified by the user in an ad-hoc manner (the range of the grid may be too narrow or the spacings between different elements of the grid may be too small). Computing the entire *solution path* $\{\widehat{\beta}^{\ell_1,\lambda}\}_{\lambda\in(0,\infty)}$ can in principle be done with the help of the lasso modification of the LARS algorithm [55, 130], which however becomes impractically slow if both n and p are large and which, in the worst case, may have exponential runtime complexity in p [102].

In summary, the lasso enjoys both favourable computational properties as well as theoretical guarantees with regard to prediction and estimation, which however are coupled to proper tuning of the regularization parameter. While the lasso takes advantage of sparsity and provides exactly sparse solutions, it often fails to achieve support recovery.

These two points imply that applying the lasso to practical problems requires some care and that there is room left for improvement, which, depending on the situation, can be filled by alternative methods.

Greedy algorithms. A third class of approaches tries to solve the ℓ_0-constrained least squares problem (1.4) in a greedy manner by incrementally building up an estimate for the support of β^*. Orthogonal matching pursuit (OMP, [103]), also known as forward selection [167], can be seen as the basic variant in this context. The main advantage of OMP is its low computational complexity: as long as the computations are properly organized, OMP is not much more expensive than solving a least squares problem restricted to the variables in S [15]. On the other hand, support recovery via OMP requires the same restrictive condition as the lasso [175]. The forward-backward algorithm proposed in [177] improves in this regard by alternating between forward and backward steps, which allows one to get rid of wrong predictor variables selected at earlier stages. On the downside, additional complications regarding the stopping criterion as well as increased computational costs are involved.

1.1.5 Estimation procedures for sparse, non-negative high-dimensional linear models and contributions of this chapter

We now turn our attention to the sparse, non-negative case with parameter set $B_0^+(s;p)$, which is at the center of interest in this thesis. Non-negativity is a particularly relevant constraint since non-negative data are frequently encountered in various areas of modern data analysis. Common examples include pixel intensities of a greyscale image, adjacency matrices, time measurements, bag-of-words or other forms of count data, power spectra or economic quantities such as prices, incomes and growth rates. In this thesis, we explore in detail to what extent the additional non-negativity constraint simplify the estimation problem. Most of the approaches mentioned in the previous subsection admit a straightforward modification accounting for non-negativity. For example, in place of the lasso, one may use the *non-negative lasso*

$$\widehat{\beta}^{\ell_1^+,\lambda} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p_+} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda \mathbf{1}^\top\beta, \;\; \lambda > 0. \tag{1.15}$$

While it turns out that under non-negativity, the non-negative lasso improves over the lasso in practice, it will be shown in this thesis that the non-negative lasso inherits the major shortcomings of its unconstrained counterpart, notably the requirement of specifying the tuning parameter λ. The fact that all popular sparse estimation techniques depend on tuning parameters, whose proper choice can be notoriously hard in practice, along with the observation that non-negativity may be a powerful constraint, motivates us to propose *non-negative least squares* (NNLS) estimation as an alternative. NNLS yields an estimator $\widehat{\beta}$ as

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p_+} \frac{1}{n}\|y - X\beta\|_2^2. \tag{1.16}$$

Both (1.15) and (1.16) involve the solution of similar convex quadratic programs which, due to the simplicity of the constraints, constitute basic problems in convex optimization for which many solvers exist that handle efficiently even problems with large

problem dimensions n and p, see [86] for an example. From a statistical perspective, at first glance, one may have considerable doubts regarding the usefulness of (1.16) in a high-dimensional, sparse setting. These doubts arise from the failure of standard least squares estimation in this setting (cf. the discussion in §1.1.2) and the fact that (1.16) is a pure fitting approach that does not seem to allow one to take advantage of sparsity. Thus the use of NNLS appears to contradict a paradigm of high-dimensional statistical inference according to which some appropriate form of regularization is necessary to deal with high dimensionality and to exploit sparsity.
On the other hand, NNLS has been used with quite some success in applications including deconvolution and unmixing problems in diverse fields such as acoustics [98], astronomical imaging [5], hyperspectral imaging [147], genomics [96], proteomics (see §1.5), spectroscopy [53] and network tomography [108]; see [35] for a survey. Moreover, NNLS is the major building block of the standard alternating optimization scheme for *non-negative matrix factorization*, a meanwhile established tool for dimension reduction of non-negative data (see §2).
The appealing empirical performance of NNLS reported in the above references has, in our opinion, not been given sufficient theoretical explanation. An early reference is [53] dating back already two decades. The authors show that, depending on X and the sparsity level, NNLS may have a 'super-resolution'-property that permits reliable estimation of β^*. Rather recently, sparse recovery of non-negative signals in the noiseless case has been studied in [17, 52, 165, 166]. One important finding of this body of work is that non-negativity constraints alone may suffice for sparse recovery, without the need to use ℓ_1-minimization. On the other hand, it remains unclear whether similar results continue to hold in a more realistic noisy setup. In this thesis, we present a thorough statistical analysis whose goal is to close this gap and to reconcile practical and theoretical performance of NNLS within a coherent theory of sparse, non-negative high-dimensional regression and signal recovery. Below, we summarize the key contributions of this chapter.

- We characterize a *self-regularizing property* which NNLS exhibits for a certain class of design matrices that turn out to be tailored to the non-negativity constraints. The self-regularizing property tends to be fulfilled in typical domains of applications of NNLS, in which both the design matrix and the observations are non-negative. As a result, we improve the understanding of the empirical success of NNLS.

- Moreover, the self-regularizing property yields an explicit link between NNLS and the non-negative lasso, which allows us to resolve the apparent conflict to existing theory of high-dimensional statistical inference. Elaborating further on that connection, we show that NNLS achieves near-optimal performance with regard to prediction and estimation in ℓ_q-norm, $q \in [1, 2]$, under a condition on X that combines the self-regularizing property with the *restricted eigenvalue condition* [11] used in the analysis of the lasso. Optimality of NNLS under this condition is given further support by deriving a lower bound on the asymptotic minimax rate of estimating a sparse, non-negative vector in ℓ_2-norm.

- Using a different set of conditions on X, we derive an upper bound for NNLS on the rate of estimation in ℓ_∞-norm and suggest hard thresholding of the NNLS

estimator to recover the support of β^*. An entirely data-driven procedure for the choice of the threshold is devised. Altogether, under appropriate conditions, NNLS is shown to be near-optimal regarding prediction, estimation and support recovery, without requiring tuning. The last aspect is seen to be a crucial advantage in practice over conventional sparse estimation procedures.

- We demonstrate the practical usefulness of NNLS in the challenging problem of feature extraction from protein mass spectra. This specific problem turns out to be rather instructive because several standard assumptions made in the analysis of sparse estimation procedures, such as constant variance of the error terms and correctness of the specified model, fail to be satisfied. We explain how existing sparse estimation procedures need be modified to perform satisfactorily in such situation.

- Fundamental to the success respectively failure of NNLS is an interesting phase transition phenomenon in high-dimensional geometry concerning the combinatorial structure of polyhedral cones, which parallels existing theory on ℓ_1/ℓ_0-equivalence [43, 46, 49, 50, 51]. Aspects of the phenomenon described in this thesis have already been discussed in prior work [17, 52, 165, 166], and we extend and unify the results therein.

1.2. Preliminaries

We here introduce a few notions required in substantial portions of our analysis.

General linear position. For $k \in \{0, \dots, p\}$, let $\mathcal{J}(k) = \{J \subseteq \{1, \dots, p\} : |J| = k\}$. We say that the columns of X are in general linear position in $\mathbb{R}^n$ if the following condition (GLP) holds

$$\text{(GLP)}: \quad \forall J \in \mathcal{J}(n \wedge p) \;\; \forall \lambda \in \mathbb{R}^{|J|} \;\; X_J \lambda = 0 \implies \lambda = 0. \tag{1.17}$$

The condition states that X does not contain more linear dependencies than it must. This can be seen as the generic case considering the fact that (GLP) holds with probability one if the columns of X are drawn independently from a probability distribution which is absolutely continuous w.r.t. the Lebesgue measure. However, verifying (GLP) for a given X is computationally not tractable in general if $p > n$. Assuming (GLP) avoids cumbersome case distinctions and hence simplifies our presentation. For this reason, we suppose throughout the chapter that (GLP) is satisfied, but it is mentioned explicitly whenever a certain property requires (GLP) to hold.

Normalization. For the analysis in the presence of noise, we assume that the columns of X are normalized such that $\|X_j\|_2^2 = n$ (for deterministic X) respectively $\mathbf{E}[\|X_j\|_2^2] = n$ (for random X), $j = 1, \dots, p$. According to the linear model (1.1), this may be assumed without loss of generality by a re-scaling of the form $(XD)(D^{-1}\beta^*)$ for some diagonal matrix D having positive diagonal elements. After fixing the scale of the columns of X, the signal-to-noise ratio of the problem only depends on the magnitude of the entries of β^* and the scale of the error terms.

Background on sub-Gaussian random variables. Sub-Gaussian random variables have the property that their tail probabilities can be bounded as for Gaussian random variables. This makes them particularly convenient for analysis. Various other properties of this class of random variables can be found in [20]. We here only mention facts that are frequently used throughout this chapter.

Definition 1.2. *Let Y be a random variable and let $Z = Y - \mathbf{E}[Y]$. We say that Y is sub-Gaussian if there exists $\sigma > 0$ so that*

$$M_Z(t) := \mathbf{E}[\exp(tZ)] \leq \exp(\sigma^2 t^2/2) \ \forall t \in \mathbb{R}.$$

The map $t \mapsto M_Z(t)$ is called the moment-generating function of Z and σ is referred to as the sub-Gaussian parameter of Y.

More generally, a random vector Y taking values in $\mathbb{R}^n$, $n \geq 1$, is called sub-Gaussian if the random variables $Y_u = \langle Y, u \rangle$ are sub-Gaussian for all $u \in \mathbb{R}^n$ having unit Euclidean norm. Note that if $Z_1, \ldots, Z_n$ are i.i.d. copies of a zero-mean sub-Gaussian random variable Z with parameter σ and $v \in \mathbb{R}^n$, then $\sum_{i=1}^n v_i Z_i$ is sub-Gaussian with parameter $\sigma \|v\|_2$. The following tail bound follows from the Chernov method (e.g. [106], §2.1).

$$\mathbf{P}(|Z| > z) \leq 2 \exp\left(-\frac{z^2}{2\sigma^2}\right), \quad z \geq 0. \tag{1.18}$$

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

$$\mathbf{P}\left(\max_{1 \leq j \leq p} |v_j^\top \mathbf{Z}| > \sigma \max_{1 \leq j \leq p} \|v_j\|_2 \left(\sqrt{2 \log p} + z\right)\right) \leq 2 \exp\left(-\frac{1}{2} z^2\right), \quad z \geq 0. \tag{1.19}$$

'With high probability'. Occasionally, we use the phrase 'with high probability', meaning 'with probability tending to one as n tends to infinity'.

1.3. Exact recovery and neighbourliness of high-dimensional polyhedral cones

This section is devoted to the exact recovery problem in the noiseless case as stated in §1.1.1. Here the goal is to recover $\beta^* \in B_0^+(s; p)$ from observations $y = X\beta^*$ in case that $n < p$, which is assumed throughout the whole section.

1.3.1 Non-negative solutions to underdetermined linear systems of equations and error correcting codes

It is clearly not possible to recover β^* from solving the linear system of equations

$$\text{find } \beta \quad \text{such that } X\beta = y, \tag{1.20}$$

because if $n < p$, that linear system is underdetermined and hence has infinitely many solutions. Alternatively, we may explicitly seek for sparse, non-negative solutions of

(1.20) by considering one of the following two problems:

$$\begin{aligned} &\text{find } \beta \in B_0^+(s;p) && \text{such that } X\beta = y, \\ \text{or } &\min_{\beta\in\mathbb{R}^p_+} \|\beta\|_0 && \text{such that } X\beta = y, \end{aligned} \tag{1.21}$$

where the first one requires s to be known. As discussed in §1.1.4, none of these two is practical for computational reasons. Alternatively, one could drop the combinatorial term in (1.21) and only require a solution of (1.20) to be non-negative. This yields the linear feasibility problem

$$(P_+): \quad \text{find } \beta \in \mathbb{R}^p_+ \quad \text{such that } X\beta = y.$$

Problem (P_+) is a special linear program for which many efficient solvers exist. However, it is a priori unclear whether the non-negativity constraints alone suffice to ensure recovery of β^*, i.e. whether it holds that

$$\mathcal{F}_{(P_+)} := \{\beta \in \mathbb{R}^p_+ : \ X\beta = y\} = \{\beta^*\}, \tag{1.22}$$

i.e. the feasible set $\mathcal{F}_{(P_+)}$ of (P_+) consists only of a single element, in which case (P_+) and (1.21) would be equivalent. Aspects of this question have been studied in prior work [17, 52, 165, 166], and the purpose of this section is to unify and extend these results within a common framework. The section is also intended to provide geometrical foundations we will build on when analysing NNLS (1.16) in §1.4.

Implications for the design of error correcting codes. The question of recoverability of β^* from (P_+) can be related to a question in the theory of error correcting codes [101]. This provides additional motivation for studying the recovery problem. A similar connection for ℓ_1-minimization (1.13) into the same direction are discussed in [30, 43, 51, 132]. Suppose one wants to transmit a message represented by $\theta^* \in \mathbb{R}^m$ in a way such that occasional transmission errors can be perfectly corrected by a receiver. This can be achieved by adding redundancy to the message. Specifically, we encode θ^* into $u^* = N\theta^*$ with $N \in \mathbb{R}^{p\times m}$ with $p = m + n$ for some positive integer n, while the receiver obtains a corrupted version $u = u^* + \beta^*$, where $\beta^* \in B_0^+(s;p)$. Given N, the goal of the receiver is to decode u to obtain the original message θ^*. Now the questions is whether successful decoding can be achieved by means of the linear feasibility problem

$$(P_+^\star): \quad \text{find } \theta \in \mathbb{R}^m \quad \text{such that } u - N\theta \in \mathbb{R}^p_+.$$

There is an equivalence between the problem of recovery via (P_+) and decoding via $(P_+^\star)$ as captured by the following statement.

Proposition 1.3. *Let $X \in \mathbb{R}^{n\times p}$ have full rank and let $N \in \mathbb{R}^{p\times m}$, $m = p - n$, be a matrix whose columns $\{N_1, \dots, N_m\}$ form a basis of $\mathcal{N}(X)$. Let further $\theta^* \in \mathbb{R}^m$, $u^* = N\theta^*$, $\beta^* \in B_0^+(s;p)$, $u = u^* + \beta^*$ and $y = X\beta^*$. Then θ^* is the unique solution of $(P_+^\star)$ if and only if β^* is the unique solution of (P_+).*

Proof. Suppose first that β^* is the unique solution of (P_+). Let $\widehat{\theta}$ be a solution of $(P_+^\star)$ and set $\widehat{\beta} = u - N\widehat{\theta} \succeq 0$. We then have

$$X\widehat{\beta} = X(u - N\widehat{\theta}) = Xu = X(u^* + \beta^*) = X(N\theta^* + \beta^*) = X\beta^*,$$

where we have used that $XN = 0$ by construction. Since β^* is the unique solution of (P_+), it follows that $\widehat{\beta} = \beta^*$. As a result, $\beta^* = u - N\widehat{\theta} \implies N\widehat{\theta} = N\theta^*$. Since the columns of N are linearly independent, $\widehat{\theta} = \theta^*$. For the opposite direction, suppose that there exists $\widehat{\beta} = \beta^* + \delta$, $0 \neq \delta \in \mathcal{N}(X)$ solving (P_+). As the columns of N constitute a basis of $\mathcal{N}(X)$, there exists $0 \neq \alpha \in \mathbb{R}^m$ such that $\delta = N\alpha$. Now set $\widehat{\theta} = \theta^* - \alpha$. We then have

$$u - N\widehat{\theta} = u - N(\theta^* - \alpha) = \beta^* + \delta = \widehat{\beta} \succeq 0,$$

i.e. $\widehat{\theta} \neq \theta^*$ is a solution of $(P_+^\star)$. □

1.3.2 Geometry of polyhedral cones

We now address the question of recoverability from a geometric point of view. For this purpose, we consider

$$\mathcal{C}_X = \{z \in \mathbb{R}^n : \ z = X\lambda, \ \ \lambda \in \mathbb{R}^p_+\} \subseteq \mathbb{R}^n, \tag{1.23}$$

the conic hull generated by the columns of X. In the following, we discuss several basic properties of $\mathcal{C}_X$ and conclude with a necessary and sufficient geometric condition for (1.22) to hold. Some of these properties are re-proved here for the sake of completeness. For more background, we refer to standard literature on convex geometry [40, 129, 183]. We may think of $y = X\beta^*$ as some point contained in $\mathcal{C}_X$ as y can be expressed as a non-negative combination of the generators $\{X_j\}_{j=1}^p$ of $\mathcal{C}_X$. In this context, the question under consideration can be rephrased as whether y happens to have a *unique* representation as a non-negative combination of $\{X_j\}_{j=1}^p$. The latter turns out to have a positive answer if and only if y is contained in the boundary of $\mathcal{C}_X$, which implies that it is rather simple to classify the elements of $\mathcal{C}_X$ according to whether or whether not they give rise to recoverability. We then relate this observation to a concise condition involving X and the support of β^*.

Interior of $\mathcal{C}_X$. The next statement yields a necessary condition for (1.22) to hold.

Proposition 1.4. *Let $X = [X_1 \ \dots \ X_p]$ have its columns in general linear position in $\mathbb{R}^n$, i.e. condition (GLP) in (1.17). Then $\mathcal{C}_X$ has non-empty interior and any $y \in \operatorname{\mathbf{int}} \mathcal{C}_X$ does **not** have a unique representation as non-negative combination of $\{X_1, \dots, X_p\}$.*

The first part of the statement is an immediate consequence of general linear position, which implies that the range of X is $\mathbb{R}^n$. In particular, for any unit vector $u \in \mathbb{R}^n$, there exist coefficients γ such that $u = X\gamma$. Now let $y = X\lambda$ for $\lambda \succ 0$. Then, there exists $t > 0$ sufficiently small such that $y + tu = X(\lambda + t\gamma) = X\lambda'$ for $\lambda' \succ 0$. Since u is arbitrary, $\mathcal{C}_X$ contains an Euclidean ball in $\mathbb{R}^n$ and thus has non-empty interior. In order to prove the second part of the proposition, we state and prove an additional lemma.

Lemma 1.5. *Let X be as in Proposition 1.4 and let $0 \neq y \in \operatorname{\mathbf{int}} \mathcal{C}_X$. Then there exist linearly independent points $\{z_1, \dots, z_n\} \subset \operatorname{\mathbf{int}} \mathcal{C}_X$ such that $y = \frac{1}{n}\sum_{j=1}^n z_j$.*

Proof. Pick $\{u_1, \ldots, u_{n-1}\} \subset \mathbb{R}^n$ so that $\{u_1, \ldots, u_{n-1}, y\}$ are linearly independent. Further let $u_n = -\sum_{j=1}^{n-1} u_j$. Now set

$$z_j = y + \alpha u_j,\ j = 1, \ldots, n, \ \text{ so that } y = \frac{1}{n}\sum_{j=1}^{n} z_j,$$

where $\alpha > 0$ is chosen such that $z_j \in \mathbf{int}\, \mathcal{C}_X$, $j = 1, \ldots, p$ (such α must exist since $y \in \mathbf{int}\, \mathcal{C}_X$). To see that the $\{z_j\}_{j=1}^n$ are linearly independent, note that for real numbers θ_j, $j = 1, \ldots, n$,

$$\sum_{j=1}^{n} \theta_j z_j = 0 \iff \alpha \sum_{j=1}^{n} \theta_j u_j + y \sum_{j=1}^{n} \theta_j = 0. \tag{1.24}$$

If $\sum_{j=1}^n \theta_j \neq 0$, then

$$y = -\frac{\alpha}{\sum_{j=1}^n \theta_j} \sum_{j=1}^{n} \theta_j u_j,$$

which is a contradiction, since y is linearly independent of $\{u_1, \ldots, u_{n-1}\}$. Considering the case $\sum_{j=1}^n \theta_j = 0$, (1.24) requires

$$\sum_{j=1}^{n} \theta_j u_j = \sum_{j=1}^{n-1} (\theta_j - \theta_n) u_j = 0.$$

By the linear independence of the $\{u_j\}_{j=1}^{n-1}$, this can be true only if $\theta_j = c$ for all j, which, together with $\sum_{j=1}^n \theta_j = 0$, implies that $c = 0$. □

We are now in position to prove the second part of Proposition 1.4.

Proof. (Second part of Proposition 1.4). Let us first consider the case $y \neq 0$. Invoking the preceding lemma, we have $y = \frac{1}{n}\sum_{j=1}^n z_j$ for $\{z_j\}_{j=1}^n \subset \mathcal{C}_X$ linearly independent. Consequently,

$$\begin{aligned} y = \frac{1}{n}\sum_{j=1}^{n} z_j &= \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{p} \beta_{jk} X_k \quad \text{for } \{\beta_{jk}\} \text{ non-negative,} \\ &= \sum_{k=1}^{p} \frac{1}{n}\sum_{j=1}^{n} \beta_{jk} X_k \\ &= \sum_{k=1}^{p} \gamma_k X_k, \quad \text{where } \gamma_k = \frac{1}{n}\sum_{j=1}^{n} \beta_{jk}. \end{aligned}$$

There must exist indices $\{i_1, \ldots, i_m\}$, $n \leq m \leq p$ so that $\gamma_{i_k} > 0$, $k = 1, \ldots, m$. In fact, if we had $m < n$, the $\{z_j\}_{j=1}^n$ could not be linearly independent. If $m > n$, there exists $0 \neq \delta = (\delta_1, \ldots, \delta_m)$ such that $\sum_{j=1}^m \delta_j X_{i_j} = 0$. As a result, there exists $t > 0$ sufficiently small so that we can re-express $y = \sum_{j=1}^m (\gamma_{i_j} + t\delta_j) X_{i_j}$ with $\gamma_{i_j} + t\delta_j > 0$, $j = 1, \ldots, m$. For the case $m = n$, since $p > n$, we may pick $i_{m+1} \in \{1, \ldots, p\} \setminus \{i_1, \ldots, i_m\}$

and choose $0 \neq \delta = (\delta_1, \ldots, \delta_{m+1})$ such that $\sum_{j=1}^{m+1} \delta_j X_{i_j} = 0$. Note that by (GLP), we must have $\delta_{m+1} \neq 0$, which implies that δ can be chosen such that $\delta_{m+1} > 0$. There exists $t > 0$ so that we can re-express $y = \sum_{j=1}^{m+1} \eta_j X_{i_j}$ for $\{\eta_j\}_{j=1}^{m+1}$ non-negative, where

$$\eta_j = \gamma_{i_j} + t\delta_j,\ j = 1, \ldots, m, \quad \eta_{m+1} = t\delta_{m+1}.$$

Let us now turn to the case $y = 0$. Note that $0 \in \mathbf{int}\, \mathcal{C}_X$ implies that there exists $t > 0$ such that $B_t := \{v \in \mathbb{R}^n : \|v\|_2 \leq t\} \subset \mathcal{C}_X$. Consequently, for any $v \in B_t$, there exists $0 \neq \lambda \in \mathbb{R}^p_+$ and $0 \neq \gamma \in \mathbb{R}^p_+$ such that $v = X\lambda$ and $(-v) = X\gamma$ and thus $0 = X(\lambda + \gamma) = X0$, i.e. 0 does not have a unique representation as non-negative combination of columns of X. □

Pointedness of $\mathcal{C}_X$. Pointedness is a notion from the theory of convex cones ([40], p.97), which is of specific importance here. One says that $\mathcal{C}_X$ is *pointed* if $\mathcal{C}_X \cap -\mathcal{C}_X = \{0\}$.

Proposition 1.6. *Under (GLP), $\mathcal{C}_X$ is pointed if and only if $\mathcal{C}_X \subsetneq \mathbb{R}^n$.*

In particular, if $\mathcal{C}_X$ is not pointed, $\mathbf{bd}\, \mathcal{C}_X = \emptyset$ and consequently (1.22) cannot hold for *any* $\beta^* \in \mathbb{R}^p_+$. In short, without pointedness, the non-negativity constraints in (P_+) become vacuous. A proof is given at the end of the paragraph. The following condition turns out to be sufficient (and also necessary under (GLP)) for the pointedness of $\mathcal{C}_X$.

Condition 1.7.

$$(\mathcal{H}): \quad \exists w \in \mathbb{R}^n \text{ and } h \in \mathbf{int}\, \mathbb{R}^p_+ \text{ such that } X^\top w = h.$$

Condition $(\mathcal{H})$ requires the columns of X be contained in the interior of a half-space containing the origin, which is then the only extreme point of $\mathcal{C}_X$. For a given X, condition $(\mathcal{H})$ can be verified by solving a linear program.

Proposition 1.8. *Under (GLP), $\mathcal{C}_X$ is pointed if and only if Condition $(\mathcal{H})$ holds.*

For the following proof, we define

$$\mathcal{P}_X = \{z \in \mathbb{R}^n : \ z = X\lambda, \ \lambda \in T^{p-1}\}, \quad \text{where } T^{p-1} = \{\lambda \in \mathbb{R}^p_+ : \ \mathbf{1}^\top \lambda = 1\}, \quad (1.25)$$

the convex polytope (or convex hull) generated by the columns of X.

Proof. (Propositions 1.6 and 1.8). As an intermediate step, we show the equivalence

$$(\mathcal{H}) \iff 0 \notin \mathcal{P}_X. \quad (1.26)$$

Suppose that $0 \notin \mathcal{P}_X$. We then have

$$\min_{\lambda \in T^{p-1}} f(\lambda) > 0, \quad \text{where } f(\lambda) = \frac{1}{2}\|X\lambda\|_2^2.$$

Let $\widehat{\lambda}$ be a minimizer of f over T^{p-1} and set $w = X\widehat{\lambda}$. Note that $f(\widehat{\lambda}) > 0$ implies that $w \neq 0$. Moreover, the first order optimality condition (cf. [9], Proposition 2.1.2) of above minimization problem implies that

$$\left\langle \nabla f(\widehat{\lambda}), \lambda - \widehat{\lambda} \right\rangle = \left\langle X^\top w, \lambda - \widehat{\lambda} \right\rangle \geq 0 \quad \text{for all } \ \lambda \in T^{p-1}.$$

As a result, we have

$$\langle w, X\lambda\rangle \geq \left\langle w, X\widehat{\lambda}\right\rangle = \|X\widehat{\lambda}\|_2^2 > 0 \quad \text{for all } \lambda \in T^{p-1},$$

which implies $(\mathcal{H})$. For the opposite direction, note that $(\mathcal{H})$ implies that $\lambda^\top X^\top w \succ 0$ and hence $X\lambda \neq 0$ for all $\lambda \in T^{p-1}$.

Given the equivalence (1.26), we will first prove Proposition 1.8. Suppose first that $(\mathcal{H})$ holds. If there exist $\lambda, \gamma \in \mathbb{R}^p_+$ so that $z = X\lambda \in \mathcal{C}_X$ and $-z = X\gamma \in \mathcal{C}_X$, we have $X(\lambda+\gamma) = 0$. Suppose that $\mathbf{1}^\top(\lambda+\gamma) > 0$. Then we have $0 = X(\lambda+\gamma)/(\mathbf{1}^\top(\lambda+\gamma))$, i.e. $0 \in \mathcal{P}_X$, which contradicts $(\mathcal{H})$. On the other hand, $\mathbf{1}^\top(\lambda+\gamma) = 0$ implies that $\lambda = \gamma = 0$ and in turn also that $z = 0$. For the opposite direction, we use contraposition. If $(\mathcal{H})$ does not hold, then there exists $\lambda \in T^{p-1}$ so that $X\lambda = 0$. Let $S = \{j : \lambda_j > 0\}$ denote the support of λ. By (GLP), $|S| \geq n+1$. Partition $S = S_1 \cup S_2$, $S_1 \cap S_2 = \emptyset$, $|S_2| \leq n$. Then

$$X_{S_1}\lambda_{S_1} + X_{S_2}\lambda_{S_2} = 0 \iff X_{S_1}\lambda_{S_1} = -X_{S_2}\lambda_{S_2}.$$

Let $z = X_{S_1}\lambda_{S_1}$. Then $-z = X_{S_2}\lambda_{S_2} \in \mathcal{C}_X$, and $z \neq 0$ (again by (GLP)). Consequently, $\mathcal{C}_X$ is not pointed.

We now turn to the proof of Proposition 1.6. It will be shown that $0 \in \mathcal{P}_X$ if and only if $\mathcal{C}_X = \mathbb{R}^n$. Along with Proposition 1.8 and (1.26), this will conclude the proof. Suppose that $0 \in \mathcal{P}_X$. Then $\mathcal{P}_X$ must also contain a Euclidean ball centered at 0. To see this, first observe that $\mathcal{P}_X$ has non-empty interior. Otherwise, the $\{X_j\}_{j=1}^p$ would be contained in an affine subspace of $\mathbb{R}^n$, i.e. there would exist $0 \neq \alpha \in \mathbb{R}^n$ and $b \in \mathbb{R}$ such that $X^\top\alpha = b\mathbf{1}$. If $b = 0$, this would mean that the columns of X are contained in a proper linear subspace of $\mathbb{R}^n$, which would contradict (GLP). On the other hand, if $b \neq 0$, condition $(\mathcal{H})$ would be satisfied, which would contradict the assumption that $0 \in \mathcal{P}_X$. Now note that because of (GLP), $X\lambda = 0$ implies that λ has $n+1$ non-zero entries and hence that 0 is an interior point of $\mathcal{P}_X$ (cf. [183], Lemma 2.9). The fact that a whole neighbourhood of zero is contained in $\mathcal{P}_X$ implies that $\mathcal{C}_X = \mathbb{R}^n$. If $\mathcal{C}_X = \mathbb{R}^n$, choose an arbitrary $0 \neq z \in \mathbb{R}^n$ such that $z = X\lambda$ and $-z = X\gamma$ for $0 \neq \lambda \in \mathbb{R}^p_+$ and $0 \neq \gamma \in \mathbb{R}^p_+$. We then have $0 = X(\lambda+\gamma)/(\mathbf{1}^\top(\lambda+\gamma)) \in \mathcal{P}_X$. □

Boundary of $\mathcal{C}_X$. It is a fundamental result in convex geometry (cf. [183], Theorem 1.3) that $\mathcal{C}_X$ as the conic hull of finitely many elements can be represented as a finite intersection of half-spaces (*polyhedral set*) containing the origin, that is there exist $\{w_k\}_{k=1}^q \subset \mathbb{R}^n$ such that

$$\mathcal{C}_X = \{z \in \mathbb{R}^n : \langle w_k, z\rangle \geq 0, \quad k = 1, \ldots, q\}, \tag{1.27}$$

and we henceforth refer to $\mathcal{C}_X$ as *polyhedral cone*. For simplicity, we suppose that q is minimal in the sense that dropping any constraint associated with one of the $\{w_k\}_{k=1}^q$ would lead to a different set. Assuming that $\mathcal{C}_X$ has non-empty interior, it can be read off from representation (1.27) that

$$\mathbf{bd}\,\mathcal{C}_X = \{z \in \mathbb{R}^n : \langle w_k, z\rangle \geq 0, \quad k = 1, \ldots, q, \text{ and } \exists l \in \{1, \ldots, q\} \text{ s.t. } \langle w_l, z\rangle = 0\}. \tag{1.28}$$

That is, $\mathbf{bd}\,\mathcal{C}_X$ is contained in a collection of hyperplanes with normal vectors w_k, $k = 1, \ldots, q$. The intersections of these hyperplanes with $\mathcal{C}_X$ are called *facets*. Using that each of element of $\mathcal{C}_X$ can be expressed as non-negative combination of $\{X_j\}_{j=1}^p$, (1.28) in turn gives rise to the representation

$$\mathbf{bd}\,\mathcal{C}_X = \bigcup_{J\in\mathcal{F}} \mathcal{C}_{X_J}, \;\; \mathcal{F} := \{J \subseteq \{1,\ldots,p\} : \; \exists w \in \mathbb{R}^n \text{ s.t. } X_J^\top w = 0 \text{ and } X_{J^c}^\top w \succ 0\}, \tag{1.29}$$

where for any $J \subseteq \{1, \ldots, p\}$, $\mathcal{C}_{X_J}$ denotes the conic hull generated by X_J, i.e.

$$\mathcal{C}_{X_J} = \left\{ z \in \mathbb{R}^n : \; z = \sum_{j=1}^{|J|} \alpha_j X_{k_j}, \;\; \alpha_j \geq 0, \, j = 1, \ldots, |J| \right\}.$$

We use the convention $X_\emptyset = 0$ and accordingly $\mathcal{C}_{X_\emptyset} = \{0\}$. Each member of the union in (1.29) is said to be a *face* of $\mathcal{C}_X$, all of which are polyhedral cones included in $\mathcal{C}_X$. Geometrically, the vector w in (1.29) is the normal vector of a *separating hyperplane* for the respective face $\mathcal{C}_{X_J}$ and $\mathcal{C}_X \setminus \mathcal{C}_{X_J}$, e.g. [16], §2.5.1.

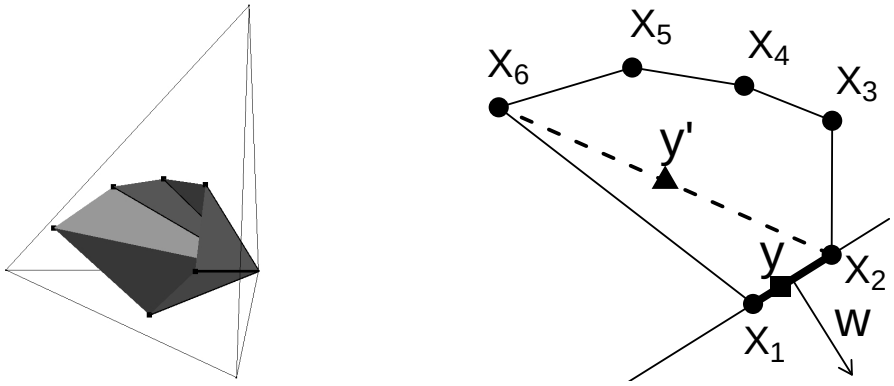


Figure 1.1: Left: Cone $\mathcal{C}_X \subset \mathbb{R}^3_+$ and $p = 6$. Right: A slice of $\mathcal{C}_X$. The columns of X are indicated by dots. The face containing y is depicted by a bold black segment which is part of a solid line representing the corresponding separating hyperplane with normal vector w. By contrast, y' has a sparse representation in terms of $\{X_2, X_6\}$, which, however, fail to form a face of $\mathcal{C}_X$.

Let us now relate the structure of $\mathbf{bd}\,\mathcal{C}_X$ to the question of recoverability according to (1.22). If $y \in \mathbf{int}\,\mathcal{C}_X$, (1.22) fails to hold in view of Proposition 1.4. On the other hand, if $y \in \mathbf{bd}\,\mathcal{C}_X$ and the columns of X are in general linear position, recoverability is ensured. In fact, under (GLP), a facet contains precisely $n-1$ linearly independent elements of $\{X_j\}_{j=1}^p$ generating the facet, and hence y can be represented as unique non-negative combination of these elements. Moreover, we gain some insight into the role of sparsity at a rough intuitive level. For $\beta^* \in \mathbb{R}^p_+$ with support S, $y \in \mathbf{bd}\,\mathcal{C}_X$ if and only if $\mathcal{C}_{X_S}$ is a face of $\mathcal{C}_X$. The smaller $|S|$, the more likely $\mathcal{C}_{X_S}$ happens to be a face. Indeed, for any $Q \subseteq S$, $\mathcal{C}_{X_Q}$ is trivially a face if $\mathcal{C}_{X_S}$ is. On the other hand, $\mathcal{C}_{X_Q}$ may be a face while $\mathcal{C}_{X_S}$ is not. We wrap up with the following statement whose proof is immediate from the above discussion.

Proposition 1.9. *Suppose that the columns of X are in general linear position and let $S \in \mathcal{J}(s)$ for $s \in \{0, 1, \ldots, n-1\}$. Then the following properties are equivalent.*

i) For any $\beta^* \in B_0^+(S;p)$, β^* is the unique solution of (P_+) with right hand side given by $y = X\beta^*$.

ii) $\mathcal{C}_{X_S}$ is a face of $\mathcal{C}_X$.

The proposition is stated for all vectors with fixed support S. Since typically only an upper bound on the cardinality of the support of β^* is available in applications, a statement asserting recoverability uniformly over the set $B_0^+(s;p)$ is desirable. This leads us to the notion of *neighbourliness* of polyhedral cones, which is defined as follows. The term neighbourliness was coined in the theory of convex polytopes to describe a corresponding phenomenon; see the discussion in §1.3.4.

Definition 1.10. *Let the columns of X be in general linear position. The polyhedral cone $\mathcal{C}_X$ is said to be s-**neighbourly**, $0 \leq s \leq n-1$, if for all $S \in \mathcal{J}(s)$, $\mathcal{C}_{X_S}$ is a face of $\mathcal{C}_X$.*

Note that zero-neighbourliness is equivalent to condition $(\mathcal{H})$. Geometrically, s-neighbourliness requires that for any $S \in \mathcal{J}(s)$, the cone $\mathcal{C}_{X_S}$ does not intersect with $\mathbf{int}\,\mathcal{C}_X$. For example, 1-neighbourliness requires that all of the $\{X_j\}_{j=1}^p$ are extreme rays of $\mathcal{C}_X$, i.e. none of them is contained in the conic hull of the others. Equipped with this notion, we state a result which eliminates the dependence of a specific support S in the previous proposition.

Proposition 1.11. *Suppose that the columns of X are in general linear position and let $s \in \{0, 1, \ldots, n-1\}$. Then the following properties are equivalent.*

i) For any $\beta^* \in B_0^+(s;p)$, β^* is the unique solution of (P_+) with right hand side given by $y = X\beta^*$.

ii) $\mathcal{C}_X$ is s-neighbourly.

1.3.3 ℓ_1/ℓ_0 equivalence and neighbourliness of polytopes

The geometric considerations above parallel a similar analysis in [49] concerning the recoverability of $\beta^* \in B_0^+(s;p)$ from observations $y = X\beta^*$ using *non-negative ℓ_1-minimization*, which is given by the linear program

$$\min_{\beta \in \mathbb{R}_+^p} \mathbf{1}^\top \beta \quad \text{such that } X\beta = y. \tag{1.30}$$

Optimization problem (1.30) can be seen as convex relaxation of (1.21) (cf. the discussion in §1.1.4). In [49], the authors provide a geometric characterization of a phenomenon which they term ℓ_1/ℓ_0 equivalence. Assuming that non-negative ℓ_0-minimization (1.21) has β^* as its unique solution, equivalence with non-negative ℓ_1-minimization amounts to

$$\operatorname*{argmin}_{\beta \in \mathcal{F}_{(P_+)}} \mathbf{1}^\top \beta = \{\beta^*\}, \tag{1.31}$$

where we recall that $\mathcal{F}_{(P_+)}$ denotes the feasible set $\{\beta \in \mathbb{R}^p_+ : X\beta = y\}$. Note that in this section, we are interested in deriving conditions under which $\mathcal{F}_{(P_+)} = \{\beta^*\}$, which trivially implies (1.31). In particular, by using non-negative ℓ_1-minimization, one can only improve over the feasibility problem (P_+) in terms of exact recovery. Conversely, if (1.31) fails to hold, then so must recovery based on (P_+). In [49], the question of whether (1.31) holds is addressed in terms of the geometry of the convex polytope

$$\mathcal{P}_{0,X} = \text{conv}\{0, X_1, \ldots, X_p\} = \{z \in \mathbb{R}^n : \ z = X\lambda, \ \lambda \in \mathbb{R}^p_+, \ \mathbf{1}^\top\lambda \leq 1\}. \tag{1.32}$$

Using a similar reasoning as above, one can show that under (GLP), $\mathcal{P}_{0,X}$ has non-empty interior, and once $p > n+1$, $y \in \textbf{int}\,\mathcal{P}_{0,X}$ implies that (1.31) must fail. The boundary of $\mathcal{P}_{0,X}$ is given by the union of its faces

$$\textbf{bd}\,\mathcal{P}_{0,X} = \bigcup_{J\in\mathcal{F}} \mathcal{P}_{0,X_J} \cup \bigcup_{J\in\mathcal{F}'} \mathcal{P}_{X_J},$$

where for $J \subseteq \{1, \ldots, p\}$, the sets $\mathcal{P}_{X_J}$ and $\mathcal{P}_{0,X_J}$ are defined analogously to $\mathcal{P}_X$ (1.25) respectively $\mathcal{P}_{0,X}$, and

$$\begin{aligned}\mathcal{F} &= \{J \subseteq \{1, \ldots, p\} : \ \exists w \in \mathbb{R}^n \text{ s.t. } X_J^\top w = 0 \text{ and } X_{J^c}^\top w \succ 0\},\\ \mathcal{F}' &= \{J \subseteq \{1, \ldots, p\} : \ \exists w \in \mathbb{R}^n \text{ and } b < 0 \text{ s.t. } X_J^\top w = b\mathbf{1} \text{ and } X_{J^c}^\top w \succ b\mathbf{1}\}.\end{aligned} \tag{1.33}$$

Note that the set $\mathcal{F}$ also indexes the faces of $\mathcal{C}_X$ (1.29) and that $(\mathcal{F} \setminus \emptyset) \subseteq \mathcal{F}'$ [2] under (GLP). In fact, under (GLP) all faces of $\mathcal{P}_{0,X}$ are simplicial. This implies that for $J \in \mathcal{F}$, $\{0\} \cup \{X_j\}_{j\in J}$ are vertices of $\mathcal{P}_{0,X}$ and that the convex hull of any subset of these vertices generates a face of $\mathcal{P}_{0,X}$; cf. Lemma 4.1 in [49]. Given that all faces of $\mathcal{P}_{0,X}$ are simplicial, one concludes that all $\beta^* \in B_0^+(S;p)$ can be recovered from observations $y = X\beta^*$ via (1.30) if $\mathcal{P}_{X_S}$ is a face of $\mathcal{P}_{0,X}$. Accordingly, all $\beta^* \in B_0^+(s;p)$ can be recovered if for all $J \in \mathcal{J}(s)$, $\mathcal{P}_{X_J}$ is a face of $\mathcal{P}_{0,X}$. In the latter case, $\mathcal{P}_{0,X}$ is called *s-outwardly neighbourly* in [49]. This has to be distinguished from classical neighbourliness employed in the theory of convex polytopes, where a polytope is called s-neighbourly if any collection of s of its vertices generate a face ([183], p.16).

1.3.4 Construction of sensing matrices

As indicated in the headline, we now look at the question of recoverability based on (P_+) from the point of view of compressed sensing (cf. the end of §1.1.3). That is, we suppose that we are free in choosing the measurement or 'sensing' matrix X, and we aim at choices that will enable us to achieve exact recovery uniformly over all $\beta^* \in B_0^+(s;p)$ for a certain level of sparsity s with as few linear measurements $y = X\beta^*$ as possible. In geometric terms, we look for matrices achieving a high level of neighbourliness for given dimensions $n < p$.

Lower bound on the number of measurements. In the theory of convex polytopes, it is known that a polytope in $\mathbb{R}^n$ cannot be more than $\lfloor n/2 \rfloor$-neighbourly, unless it is a simplex ([183], p.16). In particular, as long as $p > n+1$, $\mathcal{P}_X$ cannot be more than

[2] Recall that we use the convention $X_\emptyset = 0$, which implies that $\emptyset \neq \mathcal{F}'$.

$\lfloor n/2 \rfloor$-neighbourly and hence $\mathcal{P}_{0,X}$ cannot be more than $\lfloor n/2 \rfloor$-outwardly neighbourly, which in turn implies that $\mathcal{C}_X$ cannot be more than $\lfloor n/2 \rfloor$-neighbourly. This shows that we need at least $n \geq 2s$ measurements for recovery based on non-negative ℓ_1-minimization or (P_+). For non-negative ℓ_1-minimization, this lower bound can be attained by taking the columns X as the vertices of the so-called trigonometric cyclic polytope, a construction due to [63]. For recovery based on (P_+), it is shown in [166] that $n \geq 2s+1$ measurements are needed. In [52] an explicit construction, the so-called low-frequency partial Fourier matrices, is given for which that bound is attained. Notably, this construction − which is again based on the trigonometric cyclic polytope − is deterministic, whereas matrices popular in compressed sensing are realizations of certain ensembles of random matrices. On the downside, recovery based on low-frequency partial Fourier matrices is numerically not stable as already submatrices formed from a small number of columns are ill-conditioned.

Null space analysis. In the following, we lay the foundation of our main results in this subsection concerning the degree of neighbourliness of polyhedral cones generated by several classes of *random* matrices, which do not suffer from the ill-conditioning issue of the low-frequency partial Fourier matrices.

Restricted nullspace condition. We fix $\beta^* \in B_0^+(S;p)$ and re-consider the question whether we have recoverability (1.22) given measurements $y = X\beta^*$. We have seen that condition $(\mathcal{H})$ (Condition 1.7) is necessary, thus we suppose in the sequel that it holds. Then, the feasibility problem (P_+) is equivalent to the problem

$$(P_{+,h}): \quad \text{find } \beta \quad \text{such that } \beta \in \mathcal{F}_{(P_+)} \text{ and } h^\top \beta = w^\top y,$$

where $h \succ 0$ is a vector such that $X^\top w = h$ according to condition $(\mathcal{H})$. To see this equivalence, observe that for any $\beta \in \mathcal{F}_{(P_+)}$, we have that $h^\top \beta = w^\top X \beta = w^\top y$. This was first noted in [17]. Now suppose that $\mathcal{F}_{(P_+)}$ is not a singleton. Then there exists $0 \neq \delta \in \mathcal{N}(X)$ such that

$$h^\top \beta^* = h^\top(\beta^* + \delta) \iff h_{S^c}^\top \delta_{S^c} = -h_S^\top \delta_S \quad \text{and} \quad \delta_{S^c} \succeq 0, \tag{1.34}$$

where $\delta_{S^c} \succeq 0$ is necessary for $\beta^* + \delta \in \mathcal{F}_{(P_+)}$ as $\beta^*_{S^c} = 0$. Property (1.34) implies that $\mathcal{N}(X)$ has a non-trivial intersection with the sets

$$\begin{aligned}
&\{\delta \in \mathbb{R}^p : \ h_{S^c}^\top \delta_{S^c} \leq -h_S^\top \delta_S \text{ and } \delta_{S^c} \succeq 0\} \\
&\subseteq \left\{\delta \in \mathbb{R}^p : \ h_{S^c}^\top \delta_{S^c} \leq \sum_{j \in S} h_j |\delta_j| \text{ and } \delta_{S^c} \succeq 0\right\} \\
&\subseteq \left\{\delta \in \mathbb{R}^p : \ \sum_{j \in S^c} h_j |\delta_j| \leq \sum_{j \in S} h_j |\delta_j|\right\} \\
&\subseteq \left\{\delta \in \mathbb{R}^p : \ \|\delta_{S^c}\|_1 \leq \frac{\max_{1 \leq j \leq p} h_j}{\min_{1 \leq j \leq p} h_j} \|\delta_S\|_1\right\} \\
&= \{\delta \in \mathbb{R}^p : \ \|\delta_{S^c}\|_1 \leq \eta_h \|\delta_S\|_1\} =: \mathcal{C}(S, \eta_h), \quad \text{where } \eta_h = \max_{1 \leq j \leq p} h_j \Big/ \min_{1 \leq j \leq p} h_j.
\end{aligned} \tag{1.35}$$

Convex cones of the form

$$\mathcal{C}(J,\alpha) = \{\delta \in \mathbb{R}^p : \ \|\delta_{J^c}\|_1 \leq \alpha \|\delta_J\|_1\}, \quad J \subseteq \{1,\ldots,p\}, \ \ \alpha \in [1,\infty) \tag{1.36}$$

play a central role in the analysis of ℓ_1-norm based sparse recovery methods [11, 32, 39, 47, 61, 122, 178]. We also define

$$\mathcal{C}(k,\alpha) = \bigcup_{J \in \mathcal{J}(k)} \mathcal{C}(J,\alpha), \quad k \in \{0,\ldots,p\}, \ \ \alpha \in [1,\infty). \tag{1.37}$$

In accordance with the terminology in [122], we state the following condition.

Condition 1.12. *Let $k \in \{0,\ldots,p\}$ and $\alpha \in [1,\infty)$. We say that X satisfies condition* $\mathsf{RN}(k,\alpha)$ *(where* RN *abbreviates 'restricted nullspace') if $\mathcal{N}(X) \cap \mathcal{C}(k,\alpha) = \{0\}$.*

Based on observation (1.35) and the relation between recoverability from (P_+) and the degree of neighbourliness of $\mathcal{C}_X$, we state the following theorem.

Theorem 1.13. *Suppose that X satisfies condition $(\mathcal{H})$, i.e. there exists $w \in \mathbb{R}^n$ such that $X^\top w = h$ with $h \succ 0$. Let $\eta_h = \max_{1\leq j\leq p} h_j / \min_{1\leq j\leq p} h_j$ and suppose further that X satisfies condition* $\mathsf{RN}(s,\eta_h)$*. Then $\mathcal{C}_X$ is s-neighbourly.*

Proof. Consider the reasoning leading to (1.35). Under condition $\mathsf{RN}(s,\eta_h)$, we have that $\mathcal{N}(X) \cap \mathcal{C}(s,\alpha) = \{0\}$. Hence for any $\beta^* \in B_0^+(s;p)$ such that $y = X\beta^*$, (1.34) cannot hold, and consequently (P_+) has a unique solution. The assertion follows in view of Proposition 1.11. □

Theorem 1.13 indicates the route that will be taken to verify whether a certain matrix X generates a polyhedral cone that is s-neighbourly.

Verify.$(\mathcal{H})$:	Verify whether X satisfies condition $(\mathcal{H})$.
Bound.η_h:	If this is the case, upper bound $\eta_h = \max_{1\leq j\leq p} h_j / \min_{1\leq j\leq p} h_j$, where $h = X^\top w \succ 0$. The smaller η_h, the easier the next step.
Verify.$\mathsf{RN}(s,\eta_h)$:	Verify whether X satisfies condition $\mathsf{RN}(s,\eta_h)$.

The main advantage of this approach is that it allows one to handle whole classes of random matrices schematically and to make use of existing results in [134] for *Verify.*$\mathsf{RN}(s,\eta_h)$. The main disadvantage is that it is no longer possible to show superiority of (P_+), which uses the additional information of non-negativity, over ℓ_1-minimization (1.13) in terms of recovery: since $\eta_h \geq 1$, $\mathsf{RN}(s,\eta_h)$ implies $\mathsf{RN}(s,1)$, which in turn implies recoverability by means of (1.13), as stated below.

Proposition 1.14. *(e.g. [34], Proposition 2.2.11) Suppose that X satisfies condition* $\mathsf{RN}(s,1)$*. Let $\beta^* \in B_0(s;p)$ arbitrary and let $y = X\beta^*$. Then*

$$\operatorname*{argmin}_{\beta\in\mathbb{R}^p:\, X\beta=y} \|\beta\|_1 = \{\beta^*\}.$$

Specific constructions. In this paragraph, we state a series of results concerning the degree of neighbourliness of polyhedral cones $\mathcal{C}_X$ with X drawn from several random matrix ensembles. The proofs of these results, which are based on the three-step scheme below Theorem 1.13, are postponed to §1.4.7.

Centrosymmetric ensemble. We say that a random vector ξ taking values in $\mathbb{R}^n$ is *centrosymmetric* if for all $\sigma \in \{-1,1\}^n$ and each Borel set A, it holds that $\mathbf{P}(\xi \in A) = \mathbf{P}(\mathrm{diag}(\sigma)\xi \in A)$. Consider the random matrix ensemble

$$\begin{aligned}\mathrm{Ens}_0(n,p) : X = [X_1 \ \ldots X_p],\ \text{with } \{X_j\}_{j=1}^p \text{ i.i.d. centrosymmetric in } \mathbb{R}^n \\ \text{and } \mathbf{P}(X \text{ satisfies (GLP)}) = 1.\end{aligned} \tag{1.38}$$

Note that the centrosymmetric ensemble includes all random matrices with i.i.d. entries from probability distributions on $\mathbb{R}$ that possess a density w.r.t. the Lebesgue measure, such as the standard Gaussian distribution, the uniform distribution on $[-1,1]$, or a (central) t-distribution. The class (1.38) was studied in prior work [52], where it is pointed out that for any X from $\mathrm{Ens}_0(n,p)$, it holds that

$$\mathbf{P}(X \text{ satisfies } (\mathcal{H})) = \mathbf{P}(0 \le B \le n-1), \tag{1.39}$$

where B follows a binomial distribution with $p-1$ trials and probability of success equal to $\frac{1}{2}$. The identity (1.39) is a classical result known as Wendel's Theorem [168]. Upper bounding (1.39) by means of Hoeffding's inequality ([106], Proposition 2.7), we find that for $p/n > 2$,

$$\mathbf{P}(X \text{ satisfies } (\mathcal{H})) \le \exp\left(-n\,(p/n-2)^2/2\right). \tag{1.40}$$

The conclusion is that matrices from Ens_0 are not suited for recovery based on (P_+) once the undersampling ratio n/p drops below 0.5, because even the necessary condition $(\mathcal{H})$ is likely not to be satisfied. This constitutes a severe limitation, and we present four constructions below that behave more favourably.

Ensemble $\mathrm{Ens}_{\mathbf{1}}$. Consider the random matrix ensemble

$$\begin{aligned}\mathrm{Ens}_{\mathbf{1}}(n,p) : X = [\mathbf{1}^\top;\ \widetilde{X}],\ \text{where the } n-1 \text{ rows of } \widetilde{X} \text{ are i.i.d.} \\ \text{zero-mean isotropic}[3]\text{sub-Gaussian random vectors in } \mathbb{R}^p.\end{aligned} \tag{1.41}$$

We have used $[A;B]$ to denote the row-wise concatenation of two matrices A and B.

Theorem 1.15. *Let X be a random matrix from $\mathrm{Ens}_{\mathbf{1}}(n,p)$. Then there exist constants $C_1, C_2, c > 0$ so that with probability at least $1 - 2\exp(-cn)$, $\mathcal{C}_X$ is s-neighbourly as long as*

$$s \le \frac{n-1}{C_1 \log\left(C_2 \frac{p}{s}\right)}. \tag{1.42}$$

Ensemble Ens_+. Consider the random matrix ensemble

$$\mathrm{Ens}_+(n,p) :\ X = (x_{ij})_{\substack{1\le i\le n,\\ 1\le j\le p}},\ \{x_{ij}\} \text{ i.i.d. from a sub-Gaussian distribution on } \mathbb{R}_+. \tag{1.43}$$

[3] A random vector Y is called isotropic if $\mathbf{E}[\langle Y,u\rangle^2] = 1$ for all unit vectors u.

Among others, the class of sub-Gaussian distributions on $\mathbb{R}_+$ encompasses the zero-truncated Gaussian distribution[4], 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 $\{0,1,\dots,K\}$, as well as any finite mixture of these distributions.

Theorem 1.16. *Let X be a random matrix from $\mathrm{Ens}_+(n,p)$. Then there exist constants $C_0, C_1, C_2, c > 0$ so that if $n > C_0 \log p$, with probability at least $1 - 3\exp(-cn)$, $\mathcal{C}_X$ is s-neighbourly as long as*

$$s \leq \frac{n}{C_1 \log\left(C_2 \frac{p}{s}\right)}. \tag{1.44}$$

Gaussian equi-correlation.

Theorem 1.17. *Let X have n i.i.d. rows drawn from a p-dimensional Gaussian distribution with zero mean and covariance matrix $\Sigma^* = (1-\rho)I_p + \rho \mathbf{1}\mathbf{1}^\top$ for some $\rho \in (0,1)$. Then, there exists constants $C_0, C_1, C_2, c > 0$ so that if $n > C_0 \log p$, with probability at least $1 - 3\exp(-cn) - 2/p$, $\mathcal{C}_X$ is s-neighbourly as long as*

$$s \leq \frac{n}{C_1 \log\left(C_2 \frac{p}{s}\right)}.$$

Drawing random vectors from a half-space. So far, we have seen that condition $(\mathcal{H})$ plays a crucial role. One may wonder whether a generic set of points $\{X_j\}_{j=1}^p$ contained in the interior of some half-space in $\mathbb{R}^n$ gives rise to a suitable measurement matrix. More specifically, we consider the following scheme.

Construction `HALFSPACE`(n,p,t)

Input: $n, p, t > 0$

Choose w uniformly at random from the unit sphere in $\mathbb{R}^n$ and initialize $j \leftarrow 0$.
while $j < p$ **do**
  Generate a standard n-dimensional Gaussian vector g.
  if $\langle w, g \rangle > t$ **then**
    $j \leftarrow j + 1$ and $X_j \leftarrow g$.
  end if
end while
return $X = [X_1 \ \dots \ X_p]$.

Theorem 1.18. *Let X be generated by* `HALFSPACE`(n,p,t) *for some constant $t > 0$. Then there exist constants $C_1, C_2, C_3, c_1, c_2 > 0$ so that with probability at least $1 - 2\exp(-c_1 n) - c_2/p$, $\mathcal{C}_X$ is s-neighbourly as long as*

$$s \leq \frac{n}{C_1 \log(C_2 \frac{p}{s}) C_3 \log^{3/2}(p)}. \tag{1.45}$$

[4] For $t \in \mathbb{R}$, we refer to the distribution of a standard Gaussian random variable Z conditional on the event $\{Z \geq t\}$ as t-truncated Gaussian distribution.

Compared to Theorems 1.15 to 1.17, the order of neighbourliness is reduced by a polylog factor. From our proof in §1.4.7, it can be seen that this factor arises as a consequence of the crude estimate leading to the last inclusion in (1.35). It is left as an open question whether the extra polylog factor is in fact necessary.

Summary. With high probability, the ensembles $\mathrm{Ens}_{\mathbf{1}}$ and Ens_{+} as well as Gaussian matrices with equi-correlated columns give rise to neighbourly cones with neighbourliness proportional to the ambient dimension n when specializing Theorems 1.15 to 1.17 to a setting in which p, s grow proportionally with n, i.e. $p = \Theta(n) = \Theta(s)$. This finding comes as a surprising high-dimensional phenomenon, because it appears to be a mismatch with intuition developed in low dimensions. For example, when randomly sampling p points from $[0,1]^3$, one would expect that once $p > 3$, there is a significant chance that one of the points would be contained in the conic hull of the others so that even one-neighbourliness would fail. Yet, the above results tell us that such intuition is highly misleading in higher dimensions, and that one can hope for much more than only one-neighbourliness. Translated to the question of recoverability, this means that if the matrix X is generated from one of the ensembles of Theorems 1.15 to 1.17, it is possible to recover sparse vectors from (P_+) with a linear fraction of sparsity, where the precise value of the fraction depends on unspecified constants. Put in another way, the number of required linear measurements n required to ensure recovery is proportional to the sparsity level, while there is only a logarithmic dependence on the dimension p of the target. From this point of view, the situation is perhaps similarly surprising as the underlying geometry: it seems to be remarkable that it is possible to perfectly reconstruct a non-negative object from highly incomplete information.
On the other hand, the picture remains somewhat incomplete: in the more generic setting of Theorem 1.18, which comes close to the necessary condition $(\mathcal{H})$ as opposed to the rather special constructions in Theorems 1.15 to 1.17, the order of neighbourliness in (1.45) falls a bit short. It is of interest whether this can be improved.
A second issue left for future research concerns the reduction to the restricted nullspace condition, which involves the nesting (1.35). This can be avoided by considering a restricted nullspace condition directly over the smallest set in (1.35). As a result, one may be in position to show that in terms of exact recovery, (P_+) improves over ℓ_1-minimization without non-negativity constraints, cf. the discussion preceding Proposition 1.14.

Contribution and relation to prior work. The finding that there exists random matrices that permit recovery based on (P_+) in a setting where n, s, p grow proportionally is not novel; earlier constructions appear in [52, 166]. In the present work, we provide entire classes of matrices, $\mathrm{Ens}_{\mathbf{1}}$ and Ens_{+}, which include the constructions therein as special cases. In addition, random Gaussian matrices with equi-correlated columns and Construction `HALFSPACE` have not been considered before in the present context. In [52], the matrix $\widetilde{X}$ in (1.41) is taken as a matrix with i.i.d. standard Gaussian entries. The authors use the fact that the resulting s-dimensional faces of $\mathcal{C}_X$, $1 \leq s \leq n-1$, are in a one-to-one relation to the $s-1$ dimensional faces of $\mathcal{P}_{\widetilde{X}}$, the convex polytope generated by the columns of $\widetilde{X}$, and are hence in position to invoke results on neigh-

bourliness properties of $\mathcal{P}_{\widetilde{X}}$ proved earlier [50, 51]. The same reasoning could be used to 'lift' the results in [1] concerning neighbourliness properties of $\mathcal{P}_{\widetilde{X}}$, where $\widetilde{X}$ has i.i.d. zero-mean sub-Gaussian entries, to neighbourliness properties of $\mathcal{C}_X$. This would yield a result rather close to Theorem 1.15, apart from the fact that we only require $\widetilde{X}$ to have isotropic rows, but not necessarily independent entries.
In [166], the following construction has been suggested: a matrix $\widetilde{X}$ is generated by drawing its entries i.i.d. from a Bernoulli distribution with parameter $\frac{1}{2}$; X is then obtained by concatenating the thus generated $\widetilde{X}$ with a row of ones as it is done for $\mathrm{Ens}_{\mathbf{1}}$ here. Since $\widetilde{X}$ is a member of Ens_+, it is actually not necessary to add a row of ones. This modification is a simple way to ensure that $(\mathcal{H})$ is satisfied, which is already the case if all entries are of the same sign.

1.4. Non-negative least squares (NNLS) for high-dimensional linear models

In the present section, we change over to the noisy case. Here, NNLS whose definition we recall from (1.16)

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p_+} \frac{1}{n} \|y - X\beta\|_2^2,$$

constitutes the counterpart to the feasibility problem (P_+), in which one looks for non-negative solutions to underdetermined linear systems of equations. Note that if (P_+) has a feasible solution, then it is equivalent to NNLS. The goal is to derive theoretical properties of the latter given the linear model $y = X\beta^* + \varepsilon$ (1.1) as discussed at the beginning of this chapter. At a basic level, the question we will address below is whether the usefulness of the non-negativity constraints in the noiseless case can be carried over to the more realistic noisy case. For simplicity, we mostly restrict ourselves to i.i.d. zero-mean sub-Gaussian errors $\varepsilon_1, \ldots, \varepsilon_n$, which is standard in the literature on high-dimensional statistics. In this setting, we study the performance of the solution(s) $\widehat{\beta}$ with regard to prediction, estimation and support recovery. Even though $\widehat{\beta}$ is in general not uniquely determined, the properties that we will derive hold uniformly for all solutions. For simplicity, we hence speak of $\widehat{\beta}$ as the NNLS estimator. In a nutshell, it turns out that the answer to the question raised above is positive, under conditions that can be seen as natural strengthenings of condition $(\mathcal{H})$ and the restricted nullspace condition of the previous section. Under the strengthened version of condition $(\mathcal{H})$, to which we refer as *self-regularizing property*, one can establish an explicit connection between NNLS and (non-negative) ℓ_1-regularized least squares, or synonymously the (non-negative) lasso (1.12),(1.15). As a result, NNLS may achieve a comparable performance. In specific situations, NNLS may outperform the (non-negative) lasso with regard to estimation in ℓ_∞-norm and support recovery after thresholding. In addition to the fact that NNLS is free of tuning parameters, this provides some motivation for the use of NNLS in practice. In this regard, the situation is notably different from the noiseless case, where non-negative ℓ_1-minimization (1.30) is always at least as good as NNLS in terms of exact recovery.

1.4.1 Prediction error: a bound for 'self-regularizing' designs

We start by investigating the prediction error $\|X\widehat{\beta}-X\beta^*\|_2^2/n$ of NNLS. As main result of this subsection, we present an upper bound that resembles the so-called slow rate bound of the lasso [6, 70];[19], p.103. In contrast to the latter, the corresponding bound for NNLS can be established only for designs having the self-regularizing property to be introduced below. 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 as the lasso, albeit no explicit regularization is employed.

Overfitting of NNLS for orthonormal design. From the discussion concerning the pointedness of the polyhedral cone $\mathcal{C}_X$ in §1.3.2, it is immediate that NNLS may perform as poorly as ordinary least squares if $\mathcal{C}_X$ is not pointed. In fact, if $p > n$ and (GLP) holds, we have $\mathcal{C}_X = \mathbb{R}^n$ once $\mathcal{C}_X$ fails to be pointed and thus $X\widehat{\beta} = \Pi_{\mathcal{C}_X}(y) = y$, where $\Pi_{\mathcal{C}_X}$ denotes the Euclidean projection on $\mathcal{C}_X$. It follows that $\|X\widehat{\beta} - X\beta^*\|_2^2/n = \|X\widehat{\beta}^{\text{ols}} - X\beta^*\|_2^2/n = \frac{1}{n}\|\varepsilon\|_2^2$. This observation justifies the concern that NNLS as a pure fitting approach is prone to overfit. While condition $(\mathcal{H})$ (Condition 1.7) suffices to ensure pointedness of $\mathcal{C}_X$, it is not sufficient to prevent NNLS from overfitting. This can be seen when considering orthonormal design, i.e. $X^\top X = nI$[5] and $y = \varepsilon$ (i.e. $\beta^* = 0$) for a standard Gaussian random vector ε. In this case, 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-Gaussian distribution[6]. 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. This means that NNLS does not qualitatively improve over ordinary least squares. In particular, once $p = \Theta(n)$, NNLS fails to be consistent in the sense that it does not hold that $\frac{1}{n}\|X\widehat{\beta}\|_2^2 = o_{\mathbf{P}}(1)$ as $n \to \infty$. Note that compared to $(\mathcal{H})$, orthonormality imposes a much stronger constraint on the geometry of $\mathcal{C}_X$, which is then required to be contained in an orthant of $\mathbb{R}_+^p$ up to an orthogonal transformation. As rendered more precisely in §1.4.6, orthonormal design turns out to be at the edge of the set of designs still leading to overfitting.

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 direct strengthening of condition $(\mathcal{H})$. For this purpose, we define

$$\tau_0 = \left\{ \max \tau : \ \exists w \in \mathbb{R}^n, \ \|w\|_2 \leq 1 \quad \text{s.t.} \quad \frac{X^\top w}{\sqrt{n}} \succeq \tau \mathbf{1} \right\}. \tag{1.46}$$

[5] Recall that we assume that the columns of X are scaled such that $\|X_j\|_2^2 = n$, $j = 1, \ldots, p$.
[6] For a standard Gaussian random variable g, the distribution of $|g|$ is called half-Gaussian.

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

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \frac{1}{n}\|X\lambda\|_2^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda, \quad \text{where } T^{p-1} = \{\lambda \in \mathbb{R}_+^p : \mathbf{1}^\top \lambda = 1\}, \tag{1.47}$$

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. [138], §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. §1.4.6).

Condition 1.19. *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 (1.46), 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 1.20 is a qualitative statement preliminary to Theorem 1.21 below and mainly serves an illustrative purpose.

Proposition 1.20. *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 1.19. We then have*

$$\min_{\beta \succeq 0} \frac{1}{n}\|\varepsilon - X\beta\|_2^2 = \min_{\beta \succeq 0} \frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2(\mathbf{1}^\top \beta)^2 + O_{\mathbf{P}}\left(\frac{1}{\sqrt{n}}\right), \tag{1.48}$$

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 (1.48) obeys $\frac{1}{n}\|X\widehat{\beta}\|_2^2 = o_{\mathbf{P}}(1)$.

Proof. Since X satisfies Condition 1.19, by (1.46), there exists a unit vector w so that

$$\frac{X^\top w}{\sqrt{n}} = h, \quad \text{where } h \succeq \tau \mathbf{1}, \tag{1.49}$$

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 (1.49). Note that by (1.49) 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 $\widehat{\beta} \in \operatorname{argmin}_{\beta\succeq 0} \frac{1}{n}\|y - X\beta\|_2^2$ 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 $\frac{1}{n}\|X^\top\varepsilon\|_\infty = o_{\mathbf{P}}(1)$. □

In Proposition 1.20, 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. (1.15)) 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 [100] 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 (1.48) 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

$$\min_{\beta\succeq 0} \frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \tau^2(\mathbf{1}^\top\beta)^2 = \min_{\beta\succeq 0} \frac{1}{n}\|\varepsilon - \widetilde{X}\beta\|_2^2 + \gamma(\tau)\mathbf{1}^\top\beta, \tag{1.50}$$

where γ is a non-negative, monotonically increasing function of τ. Proposition 1.20 in conjunction with (1.50) 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.

Slow rate bound. The self-regularizing property of Condition 1.19 gives rise to the following general bound on the ℓ_2-prediction error of NNLS. Note that in Theorem 1.21 below, it is not assumed that the linear model is specified correctly. Instead, we only assume that there is a fixed target $\mathsf{f} = (\mathsf{f}_1, \dots, \mathsf{f}_n)^\top$ to be approximated by a non-negative combination of the columns of X.

Theorem 1.21. *Let $y = \mathsf{f} + \varepsilon$, where $\mathsf{f} \in \mathbb{R}^n$ is fixed and ε has i.i.d. zero-mean sub-Gaussian entries with parameter σ. Define*

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

Suppose that X satisfies Condition 1.19. Then, for any minimizer $\widehat{\beta}$ of the NNLS problem (1.16) and any $M \geq 0$, it holds with probability no less than $1 - 2p^{-M^2}$ that

$$\widehat{\mathcal{E}} \leq \mathcal{E}^0 + \left(\frac{6\|\beta^0\|_1 + 8\sqrt{\mathcal{E}^0}}{\tau^2} \right) (1+M)\sigma \sqrt{\frac{2\log p}{n}} + \frac{16(1+M)^2\sigma^2}{\tau^2} \frac{\log p}{n}, \tag{1.51}$$

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

Proof. Since $\widehat{\beta}$ is a minimizer of the NNLS problem and since β^0 is a feasible solution, we have that

$$\begin{aligned}
&\frac{1}{n}\|y - X\widehat{\beta}\|_2^2 \leq \frac{1}{n}\|y - X\beta^0\|_2^2 \\
\Leftrightarrow\ &\frac{1}{n}\|(\mathsf{f} + \varepsilon - X\beta^0) + X\beta^0 - X\widehat{\beta}\|_2^2 \leq \frac{1}{n}\|\mathsf{f} + \varepsilon - X\beta^0\|_2^2 \\
\Rightarrow\ &\frac{1}{n}\|X\beta^0 - X\widehat{\beta}\|_2^2 + \frac{2}{n}(\mathsf{f} + \varepsilon - X\beta^0)^\top X(\beta^0 - \widehat{\beta}) \leq 0 \\
\Rightarrow\ &\frac{1}{n}\|X\beta^0 - X\widehat{\beta}\|_2^2 \leq \frac{2}{n}(\mathsf{f} - X\beta^0)^\top X(\widehat{\beta} - \beta^0) + \frac{2}{n}\varepsilon^\top X(\widehat{\beta} - \beta^0).
\end{aligned} \tag{1.52}$$

Write $\widehat{\delta} = \widehat{\beta} - \beta^0$, $P = \{j : \widehat{\delta}_j \geq 0\}$ 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 Condition 1.19 along with (1.47).

$$\begin{aligned}
\frac{1}{n}\|X\widehat{\delta}\|_2^2 &= \widehat{\delta}_P^\top \Sigma_{PP} \widehat{\delta}_P + 2\widehat{\delta}_P^\top \Sigma_{PN} \widehat{\delta}_N + \widehat{\delta}_N^\top \Sigma_{NN} \widehat{\delta}_N \\
&\geq \tau^2 (\mathbf{1}^\top \widehat{\delta}_P)^2 - 2\|\widehat{\delta}_P\|_1 \|\widehat{\delta}_N\|_1.
\end{aligned} \tag{1.53}$$

Second, we bound the r.h.s. of (1.52). We set

$$A = \max_{1 \leq j \leq p} \left| \frac{1}{n} X_j^\top \varepsilon \right| \tag{1.54}$$

and use the bound

$$\max_{1 \leq j \leq p} \left| \frac{1}{n} X_j^\top (\mathsf{f} - X\beta^0) \right| \leq \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \|X_j\|_2 \sqrt{\frac{1}{n}\|\mathsf{f} - X\beta^0\|_2^2} = \sqrt{\mathcal{E}^0},$$

obtaining that

$$\frac{1}{n}\|X\widehat{\delta}\|_2^2 \leq 2(A+\sqrt{\mathcal{E}^0})\|\widehat{\delta}\|_1 \tag{1.55}$$

Inserting the lower bound (1.53) into (1.55), we obtain

$$\tau^2\|\widehat{\delta}_P\|_1^2 - 2\|\widehat{\delta}_P\|_1\|\widehat{\delta}_N\|_1 \leq 2(A+\sqrt{\mathcal{E}^0})(\|\widehat{\delta}_P\|_1 + \|\widehat{\delta}_N\|_1). \tag{1.56}$$

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

$$\|\widehat{\delta}_P\|_1 \leq \frac{4(A+\sqrt{\mathcal{E}^0}) + 2\|\widehat{\delta}_N\|_1}{\tau^2}, \tag{1.57}$$

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 (1.57), since $0 < \tau^2 \leq 1$). We now substitute (1.57) back into (1.52) and add $\mathcal{E}^0 = \frac{1}{n}\|X\beta^0 - \mathsf{f}\|_2^2$ to both sides of the inequality and re-arrange terms in order to obtain

$$\begin{aligned}
\widehat{\mathcal{E}} = \frac{1}{n}\|X\widehat{\beta} - \mathsf{f}\|_2^2 &\leq \mathcal{E}^0 + 2A(\|\widehat{\delta}_P\|_1 + \|\widehat{\delta}_N\|_1) \\
&\leq \mathcal{E}^0 + 2A\left(\frac{4(A+\sqrt{\mathcal{E}^0}) + 2\|\widehat{\delta}_N\|_1}{\tau^2} + \|\widehat{\delta}_N\|_1\right) \\
&\leq \mathcal{E}^0 + \frac{6A\|\beta^0\|_1 + 8(A^2 + A\sqrt{\mathcal{E}^0})}{\tau^2},
\end{aligned}$$

noting that by feasibility of $\widehat{\beta}$, one has $\widehat{\delta} \succeq -\beta^0$ and hence $\|\widehat{\delta}_N\|_1 \leq \|\beta^0\|_1$. We now control the random term A (1.54). Using (1.19) with $\mathbf{Z} = \varepsilon$, $v_j = X_j/n$, $j = 1, \ldots, p$, and $z = M\sqrt{2\log p}$, the event $\left\{A \leq (1+M)\sigma\sqrt{\frac{2\log p}{n}}\right\}$ holds with probability no less than $1 - 2p^{-M^2}$. The result follows. □

Theorem 1.21 bounds the excess error by a term of order $O(\|\beta^0\|_1\sqrt{\log(p)/n})$, 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 [70] who coined the notion of 'persistence' as distinguished from classical consistency with a fixed number of predictors. The excess error bound of Theorem 1.21 is of the same order of magnitude as the corresponding bound of the lasso ([6, 70]; [19], Corollary 6.1) that is typically referred to as slow rate bound. Since the bound (1.51) depends on τ, it is recommended to solve the quadratic program in (1.47) before applying NNLS, which is roughly of the same computational cost. Unlike Theorem 1.21, the slow rate bound of the lasso does not require any conditions on the design and is more favourable than (1.51) regarding the constants. In [75, 158], 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. In §1.1.5, NNLS has been motivated as a tool for non-negative data. In this context, the assumption of zero-mean noise in Theorem 1.21 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 1.21 continues to be applicable.

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

Within this subsection, 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 [11] to establish near-optimal performance guarantees of the lasso and the related Dantzig selector [32] with respect to estimation of β^* and and prediction under a sparsity scenario. Similar results are shown e.g. in [22, 28, 32, 110, 156, 173, 176], and we shall state results of that flavour for NNLS below. For the rest of the analysis of NNLS, the data-generating model (1.1) is considered for $\beta^* \in \mathbb{R}^p_+$ with support $S = \{j : \beta^*_j > 0\}$, $1 \leq |S| = s < n$ unless stated otherwise.

Reduction to the scheme used for the lasso. As stated in the next lemma, if the design satisfies Condition 1.19, 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

$$\mathcal{C}(S, 3/\tau^2) = \{\delta \in \mathbb{R}^p : \|\delta_{S^c}\|_1 \leq (3/\tau^2)\|\delta_S\|_1\}. \tag{1.58}$$

Recall that cones of the form $\mathcal{C}(S, \alpha)$, $\alpha \in [1, \infty)$, have already been encountered in the analysis of the noiseless case, cf. (1.35) and (1.36).

Lemma 1.22. *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 1.19. 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 $1 - 2p^{-M^2}$:*

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

Lemma 1.22 will be proved along with Theorem 1.24 below.
Under the conditions of the above lemma (Condition 1.19 is not required) and appropriate choice of λ, the lasso estimator $\widehat{\beta}^{\ell_1,\lambda}$ (1.12) has the property that $\widehat{\beta}^{\ell_1,\lambda} - \beta^* \in \mathcal{C}(S, 3)$ with high probability, which is crucial to the analysis of the lasso in [11]; the specific value of the constant $\alpha = 3$ is not essential. Consequently, in the situation that (1.58) holds, we are in position to carry over the analysis in [11] to obtain comparable performance bounds for NNLS. In this vein, we state the following condition from [11], based on which near-optimal rates with regard to estimation and prediction can be established. This condition constitutes a direct strengthening of the restricted nullspace condition (Condition 1.12) employed in the noiseless case.

Condition 1.23. *For $k \in \{1, \dots, p\}$ and $\alpha \in [1, \infty)$, consider the sets $\mathcal{C}(J, \alpha)$, $J \in \mathcal{J}(k)$ as defined in (1.36). We say that the design satisfies condition* $\mathsf{RE}(k, \alpha)$ *(where* RE *abbreviates 'restricted eigenvalue') if there exists a constant $\phi(k, \alpha)$ so that*

$$\min_{J \in \mathcal{J}(k)} \min_{\delta \in \mathcal{C}(J,\alpha) \setminus 0} \frac{\delta^\top \Sigma \delta}{\|\delta_J\|_2^2} \geq \phi(k, \alpha) > 0. \tag{1.59}$$

Note that the corresponding restricted nullspace condition $\mathsf{RN}(k, \alpha)$ only requires the quotient in (1.59) to be positive, but not necessarily lower bounded by a constant. In this sense, condition $\mathsf{RE}(k, \alpha)$ strengthens $\mathsf{RN}(k, \alpha)$ similarly as the self-regularizing property strengthens condition $(\mathcal{H})$. Using Lemma 1.22 and Condition 1.23, the next statement and its proof follow along the lines of the analysis in [11], cf. Theorem 7.2 therein.

Theorem 1.24. *In addition to the conditions of Lemma 1.22, assume further that X satisfies condition* $\mathsf{RE}(s, 3/\tau^2)$*. It then holds for any $q \in [1, 2]$ and any $M \geq 0$ that*

$$\|\widehat{\beta} - \beta^*\|_q^q \leq \frac{2^{3q-2}}{\{\phi(s, 3/\tau^2)\}^q} \left(1 + \frac{3}{\tau^2}\right)^{2q} s \left((1 + M)\sigma \sqrt{\frac{2 \log p}{n}}\right)^q \tag{1.60}$$

$$\frac{1}{n}\|X\widehat{\beta} - X\beta^*\|_2^2 \leq \frac{8(1 + M)^2\sigma^2}{\phi(s, 3/\tau^2)} \left(1 + \frac{3}{\tau^2}\right)^2 \frac{s \log p}{n}, \tag{1.61}$$

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

Proof. (Lemma 1.22 and Theorem 1.24). We start with the proof of Lemma 1.22, building on ideas already used in the proof of Theorem 1.21, where we replace β^0 by β^* as well as f by $X\beta^*$, and note that $\mathcal{E}^0 = 0$. Let $P = \{j : \widehat{\delta}_j \geq 0\}$ and $N = \{j : \widehat{\delta}_j < 0\}$. First note that $S^c \subseteq P$ and $N \subseteq S$. Hence, we obtain the following analog to (1.56).

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

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 $\widehat{\delta} \in \mathcal{C}(S, 3/\tau^2)$ as in the left event of Lemma 1.22 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 left event of Lemma 1.22 occurs. Otherwise, we conclude that the second event of the lemma occurs by controlling A as in the proof of Theorem 1.21. Let us now turn to the proof of Theorem 1.24 conditional on the left event of Lemma 1.22, i.e. $\{\widehat{\delta} \in \mathcal{C}(S, 3/\tau^2)\}$. We may thus invoke condition (1.59), which, when applied to (1.52), yields

$$\phi\left(s, 3/\tau^2\right)\|\widehat{\delta}_S\|_2^2 \leq \frac{1}{n}\|X\widehat{\delta}\|_2^2 \leq 2A(\|\widehat{\delta}_S\|_1 + \|\widehat{\delta}_{S^c}\|_1) \leq 2\left(1 + 3/\tau^2\right)A\|\widehat{\delta}_S\|_1, \tag{1.62}$$

where for the rightmost inequality, we have used that $\widehat{\delta} \in \mathcal{C}(S, 3/\tau^2)$. It follows that

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

The preceding bound in turn implies

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

Controlling A as at the end of the proof of Theorem 1.21, the previous two displays yield (1.60) for $q=1$ and the bound (1.61) on the prediction error. We now bound $\|\widehat{\delta}\|_2$. Let $U = S \cup T$, where $T \subseteq S^c$ denotes the index set corresponding to the s largest components (in absolute value) of $\widehat{\delta}$ outside S. Likewise, let V denote the index set of the overall s largest components of $\widehat{\delta}$. First note that since $\widehat{\delta} \in \mathcal{C}(S, 3/\tau^2)$, it also holds that $\widehat{\delta} \in \mathcal{C}(V, 3/\tau^2)$. Invoking the restricted eigenvalue condition, we have that

$$\frac{\frac{1}{n}\|X\widehat{\delta}\|_2^2}{\|\widehat{\delta}_U\|_2^2} \geq \frac{\frac{1}{n}\|X\widehat{\delta}\|_2^2}{2\|\widehat{\delta}_V\|_2^2} \geq \frac{1}{2}\phi(s, 3/\tau^2) \implies \frac{1}{n}\|X\widehat{\delta}\|_2^2 \geq \frac{1}{2}\phi(s, 3/\tau^2)\|\widehat{\delta}_U\|_2^2 \tag{1.63}$$

In the sequel, we bound $\|\widehat{\delta}_{U^c}\|_2$ in terms of $\|\widehat{\delta}_U\|_2$. Noting that for any $v \in \mathbb{R}^d$, the j-th largest (in absolute value) element satisfies $|v|_{(j)} \leq \|v\|_1/j$, $j = 1, \ldots, d$, we obtain

$$\|\widehat{\delta}_{U^c}\|_2^2 \leq \|\widehat{\delta}_{S^c}\|_1^2 \sum_{j=s+1}^{\infty} \frac{1}{j^2} \leq \frac{1}{s}\|\widehat{\delta}_{S^c}\|_1^2.$$

For the first inequality, we bound the j-th largest element $|\widehat{\delta}_{U^c}|_{(j)}$ of $\widehat{\delta}_{U^c}$ as $|\widehat{\delta}_{U^c}|_{(j)} \leq \|\widehat{\delta}_{S^c}\|_1/(s+j)$, $j = 1, \ldots, p-2s$, recalling that the s largest components (in absolute value) of $\widehat{\delta}_{S^c}$ are contained in $\widehat{\delta}_T$. Consequently,

$$\|\widehat{\delta}_{U^c}\|_2 \leq \frac{1}{\sqrt{s}}\|\widehat{\delta}_{S^c}\|_1 \leq \frac{1}{\sqrt{s}}\frac{3}{\tau^2}\|\widehat{\delta}_S\|_1 \leq \frac{3}{\tau^2}\|\widehat{\delta}_S\|_2 \leq \frac{3}{\tau^2}\|\widehat{\delta}_U\|_2,$$

where for the second inequality, we have again used that $\widehat{\delta} \in \mathcal{C}(S, 3/\tau^2)$. In total, we obtain

$$\|\widehat{\delta}\|_2 \leq (1+3/\tau^2)\|\widehat{\delta}_U\|_2. \tag{1.64}$$

On the other hand, from (1.62), we obtain the upper bound

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

Combining the previous bound with (1.63), we obtain

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

which, when substituted into (1.64), yields

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

With the usual control of the term A, we obtain (1.60) for $q=2$. The general ℓ_q-bound results from the inequality $\|\widehat{\delta}\|_q^q \leq \|\widehat{\delta}\|_1^{2q-1}\|\widehat{\delta}\|_2^{2(q-1)}$, which holds for all $q \in [1,2]$. So far, we have proved that the assertion of Theorem 1.21 holds conditional on the left event of Lemma 1.22. Conditional on the right event, we immediately deduce (1.60) (noting that $\tau^2 \leq 1$) and thus as well (1.61) via the upper bound on $\frac{1}{n}\|X\widehat{\delta}\|_2^2$ in terms of $\|\widehat{\delta}\|_1$ given in (1.62). □

Both Theorem 1.21 and Theorem 1.24 provide bounds on the prediction error. A noticeable difference is the dependence of the bounds on n, which is $1/\sqrt{n}$ for the former, whereas it is $1/n$ for the latter. Accordingly, one speaks of a slow respectively fast rate bound. Furthermore, when specializing Theorem 1.21 to the setting in which the model is correctly specified ($\mathcal{E}^0 = 0$), the bound on the prediction error depends on $\|\beta^*\|_1$ which may be rather unfavourable relative to the fast rate bound in a sparse regime.

It is instructive to compare the performance bounds in Theorem 1.24 with those of ℓ_0-constrained estimation in Proposition 1.1 with A substituted by $(1+M)\sigma\sqrt{2\log(p)/n}$. The bounds (1.6) on the one hand and the bounds (1.60),(1.61) on the other hand agree qualitatively, i.e. up to multiplicative constants. This is a remarkable result, since under the stated conditions, NNLS achieves performance bounds in prediction and estimation in ℓ_q-norm, $q \in [1,2]$, comparable to those of an estimation procedure which is computationally not tractable and which requires the underlying sparsity level to be known. Moreover, according to the discussion below Proposition 1.1, these performance bounds match, apart from a logarithmic factor, those of an oracle knowing the support of the target β^*.

It is worthwhile to have a closer look at the constants entering the bounds. The bounds (1.6) of ℓ_0-constrained estimation involves the quantity $\phi_{\min}(2s)$ (1.5), which is the smallest eigenvalue over all $2s \times 2s$ principal submatrices of Σ. Roughly speaking, this quantity tends to be better behaved than the 'restricted eigenvalue' $\phi(s,1)$ and hence also than $\phi(s, 3/\tau^2)$ in view of the containment $B_0(2s;p) \subseteq \cup_{J\in\mathcal{J}(s)}\mathcal{C}(J,1)$. Indeed, for $\delta \in B_0(2s;p)$, let J denote the set of its s largest components (in absolute value) so that $\|\delta_{J^c}\|_1 \leq \|\delta_J\|_1$ and thus $\delta \in \mathcal{C}(J,1)$, cf. [11], p.1710. In particular, it follows that $\phi(s,1) > 0$ implies that $\phi_{\min}(2s) > 0$.

The bounds for NNLS additionally depend on the constant τ, which quantifies the amount of self-regularization induced by the design. This is a peculiarity of NNLS relative to methods based on explicit regularization. The lasso (1.12) attains the bounds of Theorem 1.21 with $\tau^2 = 1$ (modulo a constant power of 2) provided the regularization parameter is proportional to $\sigma\sqrt{\log(p)/n}$ (cf. [11], Theorem 7.2). In summary, the conditions on the design required by NNLS are more restrictive than those of ℓ_0-constrained estimation and the lasso, and the constants in the bound are less favourable. On the positive side, NNLS achieves these bounds without the necessity of tuning or explicit knowledge of problem-specific quantities such as the sparsity level or the sub-Gaussian parameter σ of the noise.

The condition on the design required in Theorem 1.24 involves a combination of the self-regularizing property and a 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 $\phi(s,\alpha) \leq 2(1-\max_{j,k,\, j\neq k} \langle X_j, X_k\rangle /n)$. We resolve this apparent contradiction in §1.4.6 by providing designs satisfying both conditions simultaneously. The use of Condition 1.23 in place of more restrictive conditions like restricted isometry properties (RIP, [32]) 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 [122].

Results similar in spirit as Theorem 1.24 are shown in the recent paper [108] by Mein-

shausen 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 the 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' [156, 157] which is used in place of Condition 1.23. 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 1.24, a lower bound on the minimum non-zero coefficient of β^* is additionally required in the corresponding results in [108].

1.4.3 Asymptotic rate minimaxity

While Theorem 1.24 asserts that the performance of NNLS can be as good as that of an oracle apart from a logarithmic factor, it still remains unclear whether there are estimators that can achieve even better rates. In particular, it cannot be decided whether ℓ_0-constrained estimation improves over NNLS, because Proposition 1.1 only yields upper bounds which match those of Theorem 1.24 up to multiplicative constants. A common criterion for assessing the optimality of estimators is *rate minimaxity* ([153], §2). In a nutshell, the idea is to derive a lower bound on the rate of the supremal risk over all problem instances of a class of interest, which holds uniformly over all possible estimators. Accordingly, an estimator achieves rate minimaxity if it satisfies an upper bound matching that lower bound. It is important to note that a lower bound on the minimax rate always depends on the chosen problem class and loss function. To avoid digressions, we start directly with the specific problem of interest and refer to [153] and the references therein for more background.

Setup. We consider the linear model (1.1) with a design matrix from the set

$$\mathcal{X} = \{X \in \mathbb{R}^{n\times p} : \ \|X_j\|_2^2 = n, \ j = 1, \dots, p\}. \tag{1.65}$$

and $\varepsilon \sim N(0, \sigma^2 I_n)$. We are interested in lower bounding the minimax risk

$$R(B_0^+(s;p); X) = \inf_{\widehat{\theta}} \sup_{\beta^* \in B_0^+(s;p)} \mathbf{E}_{X,\beta^*}[\|\widehat{\theta} - \beta^*\|_2^2], \tag{1.66}$$

where the infimum is over all estimators $\widehat{\theta} = \widehat{\theta}(X, y)$ and $\mathbf{E}_{X,\beta^*}$ is a shorthand for $\mathbf{E}_{y\sim N(X\beta^*,\sigma^2 I_n)}$. In plain words, we study the question of how well a sparse, non-negative vector can be estimated in ℓ_2-norm from certain linear observations perturbed by additive Gaussian noise.

Prior work. Similar setups without non-negativity constraints, i.e. with parameter set $B_0(s;p)$, have been studied earlier. The lower bounds in [123] depend on specific properties of the design matrix X. This is avoided in [171] and [27]. In addition, the results in [27] are non-asymptotic and hold for any scaling of n, s and p. On the other hand, the approach taken in [27] seems to require major modifications in case

the parameter set is changed to $B_0^+(s;p)$. This is unlike the approach in [171], in which a lower bound on (1.66) results as byproduct. However, the derivations therein are sketchy, with several important details omitted. In the sequel, we fill these gaps.

Lower bound on the minimax risk. We here state the final result of this subsection and discuss its implications regarding the optimality of NNLS.

Theorem 1.25. *For the setup as given above, for any $X \in \mathcal{X}$, it holds that*

$$R(B_0^+(s;p);X) \geq (1+o(1))2\sigma^2 \frac{s}{n} \log(p/s)$$

as $p/s \longrightarrow \infty$ and $s \longrightarrow \infty$.

Before turning to the proof, let us briefly comment on this result. The lower bound equals the asymptotic minimax risk in the case of orthonormal design ([80], Theorem 8.10), which holds irrespectively of whether the parameter set is $B_0^+(s;p)$ or $B_0(s;p)$. It is intuitively appealing that for any choice of $X \in \mathcal{X}$, it is not possible to improve over the case of orthonormal design, in particular if $n < p$, and the theorem confirms that intuition. It is important to note that the result is asymptotic and does not cover arbitrary regimes of s and p. The requirement $p/s \to \infty$ does not constitute a serious restriction, since all upper bounds discussed so far, including that of the oracle estimator (1.10), require $n/s \to \infty$ for the estimation error to vanish as $n \to \infty$. On the other hand, the requirement $s \to \infty$ excludes the scaling $s = O(1)$.
In a regime where $\log(p/s) = \Omega(\log p)$, e.g. fractional power sparsity with $s = p^\nu$ for $\nu \in (0,1)$, Theorem 1.25 establishes the rate minimaxity of NNLS for the ℓ_2-error in estimation under the conditions of Theorem 1.24.

Proof of Theorem 1.25: high-level outline. As mentioned above, we follow the route in [171], which in turn builds on a technique developed in [80]. This technique hinges on the fact that the minimax risk can be lower bounded by the Bayes risk under any prior distribution supported on the parameter set. Formally, let $\mathbf{p}$ be a distribution on $\mathbb{R}_+^p$ such that $\mathbf{P}_{\beta^* \sim \mathbf{p}}(\beta^* \in B_0^+(s;p)) = 1$, we then have

$$\begin{aligned} R(B_0^+(s;p);X) &= \inf_{\widehat{\theta}} \sup_{\beta^* \in B_0^+(s;p)} \mathbf{E}_{X,\beta^*}[\|\widehat{\theta} - \beta^*\|_2^2] \\ &\geq \inf_{\widehat{\theta}} \mathbf{E}_{\mathbf{p}} \mathbf{E}_{X,\beta^*}[\|\widehat{\theta} - \beta^*\|_2^2] =: B(\mathbf{p}), \end{aligned} \tag{1.67}$$

where $B(\mathbf{p})$ is called the Bayes risk under prior $B(\mathbf{p})$ and $\mathbf{E}_{\mathbf{p}}$ is a shorthand for $\mathbf{E}_{\beta^* \sim \mathbf{p}}$. For our purpose, however, we need to consider a prior that is *not* supported on $B_0^+(s;p)$, which requires a bit of extra work. More specifically, we consider a prior of the form $\pi = \pi_{\alpha,\mu} = \prod_{j=1}^p \nu_{\alpha,\mu}$ on $\mathbb{R}_+^p$, where $\nu = \nu_{\alpha,\mu}$ is a two-point prior on $\mathbb{R}_+$ which assigns measure $\alpha \in (0,1)$ to $\mu > 0$ and measure $1-\alpha$ to 0, that is

$$\mathbf{P}_\nu(\{\mu\}) = \alpha, \quad \mathbf{P}_\nu(\{0\}) = 1 - \alpha.$$

It is clear that $\mathbf{P}_{\beta^* \sim \pi}(\beta^* \in B_0^+(s;p)) < 1$, and hence relation (1.67) does not apply to π. As a workaround, we make use of a strategy described in detail in [80], §4.11 and §8.8. The basic steps are as follows.

1. Denote by $\pi_{|s}$ the conditional distribution of $\beta^* \sim \pi$ under the event $\{\beta^* \in B_0^+(s;p)\}$ and by $B(\pi_{|s})$ the associated Bayes risk. One then bounds

$$B(\pi) \leq B(\pi_{|s}) + \varrho \leq R(B_0^+(s;p);X) + \varrho,$$

where ϱ is a remainder term. Choosing the parameter α of π as $\alpha = \alpha_{p,s,\varepsilon} = (1-\varepsilon)s/p$ and letting $s \to \infty$, $\varepsilon \to 0^+$, it will be shown that $\varrho = o(B(\pi))$. This finally yields $R(B_0^+(s;p);X) \geq (1+o(1))B(\pi)$.

2. Following [171], we lower bound $B(\pi)$ by a reduction to the case of orthonormal design, which is well studied e.g. in [80].

Reduction to the case of orthonormal design. We start with the second step. The reduction we are aiming at is based on the following observation. For $j = 1,\dots,p$, consider estimators $\widehat{\psi}_j = \widehat{\psi}_j(X, y, \{\beta_k^*\}_{k\neq j})$ for β_j^* given X and y and all regression coefficients $\{\beta_k^*\}_{k\neq j}$ except for the j-th one.

Lemma 1.26. *For $j = 1,\dots,p$, given $X, y, \{\beta_k^*\}_{k\neq j}$, the statistic*

$$z_j = \frac{1}{n} X_j^\top \left(y - \sum_{k\neq j} X_k \beta_k^* \right) = \frac{1}{n} \left(\|X_j\|_2^2 \beta_j^* + X_j^\top \varepsilon \right)$$

is sufficient [7] *for β_j^*.*

Proof. The likelihood function of β_j^* given $y, X, \{\beta_k^*\}_{k\neq j}$ is given by the expression

$$\begin{aligned}
&(2\pi)^{-n/2} \exp\left(-\frac{\|y - X\beta^*\|_2^2}{2\sigma^2} \right) \\
&= (2\pi)^{-n/2} \exp\left(-\frac{\left\| y - \sum_{k\neq j} X_k \beta_k^* \right\|_2^2}{2\sigma^2} \right) \times \left(\exp\left(-\frac{\left\| X_j \beta_j^* \right\|_2^2}{2\sigma^2} \right) \cdot \exp\left(\frac{n\beta_j^* z_j}{\sigma^2} \right) \right)
\end{aligned}$$

Hence by the factorization criterion ([94], Theorem 6.5 in §1), z_j is sufficient for β_j^*. □

Hence when considering estimators for β_j^* given $X, y, \{\beta_k^*\}_{k\neq j}$, it suffices to consider estimators of the form $\widehat{\psi}_j(z_j)$, $j = 1,\dots,p$. Note that $z_j \sim N(\beta_j^*, \sigma_n^2)$, where $\sigma_n = \sigma/\sqrt{n}$, $j = 1,\dots,p$. These observations will be used shortly. In the following, we lower bound the Bayes risk for $\beta^* \sim \pi$, equivalently $\{\beta_j^*\}_{j=1}^p \overset{\text{i.i.d.}}{\sim} \nu$ with π and ν as introduced

[7] For a definition and background, see [94], §1.6

in the preceding paragraph. Using $\mathbf{E}_j$ as a shorthand for $\mathbf{E}_{\beta_j^*\sim\nu}$, $j = 1, \dots, p$, we have

$$\begin{aligned}
B(\pi) &= \inf_{\widehat{\theta}} \mathbf{E}_\pi \mathbf{E}_{X,\beta^*}[\|\widehat{\theta} - \beta^*\|_2^2] \\
&= \inf_{\widehat{\theta}} \mathbf{E}_{2,\dots,p}\mathbf{E}_1\left[\mathbf{E}_{X,\beta^*}[\|\widehat{\theta} - \beta^*\|_2^2] \,\middle|\, \beta_2^*, \dots, \beta_p^*\right], && (1.68) \\
&= \inf_{\widehat{\theta}} \mathbf{E}_{2,\dots,p}\mathbf{E}_1\left[\mathbf{E}_{X,\beta^*}\left[\sum_{j=2}^p (\widehat{\theta}_j - \beta_j^*)^2 + (\widehat{\theta}_1 - \beta_1^*)^2\right] \,\middle|\, \beta_2^*, \dots, \beta_p^*\right] \\
&\geq \inf_{\{\widehat{\theta}_j\}_{j=2}^p} \mathbf{E}_{2,\dots,p}\mathbf{E}_1\left[\mathbf{E}_{X,\beta^*}\left[\sum_{j=2}^p (\widehat{\theta}_j - \beta_j^*)^2\right] \,\middle|\, \beta_2^*, \dots, \beta_p^*\right] + && (1.69) \\
&\quad + \inf_{\widehat{\theta}_1} \mathbf{E}_{2,\dots,p}\mathbf{E}_1\left[\mathbf{E}_{X,\beta^*}[(\widehat{\theta}_1 - \beta_1^*)^2] \,\middle|\, \beta_2^*, \dots, \beta_p^*\right] && (1.70)
\end{aligned}$$

We now consider the second term separately.

$$\begin{aligned}
&\inf_{\widehat{\theta}_1} \mathbf{E}_{2,\dots,p}\mathbf{E}_1\left[\mathbf{E}_{X,\beta^*}[(\widehat{\theta}_1 - \beta_1^*)^2] \,\middle|\, \beta_2^*, \dots, \beta_p^*\right] \\
&\geq \mathbf{E}_{2,\dots,p} \inf_{\widehat{\theta}_1} \mathbf{E}_1\left[\mathbf{E}_{X,\beta^*}[(\widehat{\theta}_1 - \beta_1^*)^2] \,\middle|\, \beta_2^*, \dots, \beta_p^*\right] \\
&\geq \mathbf{E}_{2,\dots,p} \inf_{\widehat{\psi}_1} \mathbf{E}_1\left[\mathbf{E}_{z_1\sim N(\beta_1^*,\sigma_n^2)}[(\widehat{\psi}_1 - \beta_1^*)^2]\right]
\end{aligned}$$

Regarding the second inequality, note that given $\beta_2^*, \dots, \beta_p^*$ we may replace the infimum over all estimators $\widehat{\theta}_1(X, y)$ by an infimum over all estimators $\widehat{\psi}_1(X, y, \{\beta_k^*\}_{k\neq 1})$, since this set includes all estimators only based on X and y. By sufficiency of z_1 for β_1^*, it is enough to consider estimators of the form $\widehat{\psi}_1(z_1)$, and we may accordingly replace the expectation $\mathbf{E}_{y\sim N(X\beta^*,\sigma^2)}$ by the expectation $\mathbf{E}_{z_1\sim N(\beta_1^*,\sigma_n^2)}$.
We now inspect the term

$$\inf_{\widehat{\psi}_1} \mathbf{E}_1\left[\mathbf{E}_{z_1\sim N(\beta_1^*,\sigma_n^2)}[(\widehat{\psi}_1 - \beta_1^*)^2]\right] = \inf_{\widehat{\psi}_1} \mathbf{E}_{\beta_1^*\sim\nu_{\alpha,\mu}}\left[\mathbf{E}_{z_1\sim N(\beta_1^*,\sigma_n^2)}[(\widehat{\psi}_1 - \beta_1^*)^2]\right]. \quad (1.71)$$

Asymptotic evaluation of that term for certain regimes of α and μ has already been considered in the literature. The following statement is from Lemma 8.11 in [80].

Theorem 1.27. *[80] Consider the prior $\nu_{\alpha,\mu}$ with*

$$\mu = \mu_\alpha = \sqrt{\lambda_\alpha^2 + a_\alpha^2} - a_\alpha, \quad \text{where } \lambda_\alpha = \sigma_n\sqrt{2\log\alpha^{-1}}, \quad \text{and } a_\alpha = \lambda_\alpha^{2\gamma} \quad (1.72)$$

for some $\gamma \in \left(0, \frac{1}{2}\right)$. Then as $\alpha \to 0+$

$$\begin{aligned}
\inf_{\widehat{\psi}_1} \mathbf{E}_{\beta_1^*\sim\nu_{\alpha,\mu}}\left[\mathbf{E}_{z_1\sim N(\beta_1^*,\sigma_n^2)}[(\widehat{\psi}_1 - \beta_1^*)^2]\right] &= (1 + o(1))\alpha\mu_\alpha^2 \\
&= (1 + o(1))\alpha\lambda_\alpha^2 \\
&= \sigma_n^2(1 + o(1))2\alpha\log\alpha^{-1}.
\end{aligned}$$

Hence, under the conditions of the theorem, in the limit $\alpha \to 0+$, term (1.70) is lower bounded as $\sigma_n^2(1+o(1))2\alpha\log\alpha^{-1}$. The remaining term (1.69) can be lower bounded by repeating the argument starting from (1.68) $p-1$ more times, sequentially conditioning on $\{\beta_1^*,\beta_3^*,\dots,\beta_p^*\},\dots,\{\beta_1^*,\beta_2^*,\dots,\beta_{p-1}^*\}$ and considering the infima over $\widehat{\theta}_2,\dots,\widehat{\theta}_p$ one by one. In this manner, we obtain the following lower bound on $B(\pi)$ if ν is as in Theorem 1.27:

$$B(\pi) = \inf_{\widehat{\theta}} \mathbf{E}_\pi \mathbf{E}_{X,\beta^*}[\|\widehat{\theta}-\beta^*\|_2^2] \geq p\cdot\sigma_n^2(1+o(1))2\alpha\log\alpha^{-1} \text{ as } \alpha\to 0+. \tag{1.73}$$

This concludes the second out of the two steps in the outline above.

Relating minimax risk and Bayes risk. We now turn to the first out of the two steps in the outline above. Recall that $\pi_{|s}$ denotes the conditional distribution of $\beta^*\sim\pi$ under the event $\{\beta^*\in B_0^+(s;p)\}$ and let $\delta_0=\delta_0(X,y)$ denote the Bayes estimator under the prior $\pi_{|s}$, that is

$$\delta_0 = \operatorname*{argmin}_{\widehat{\theta}} \mathbf{E}_{\pi|s}\mathbf{E}_{X,\beta^*}[\|\widehat{\theta}-\beta^*\|_2^2]. \tag{1.74}$$

We may assume that $\|\delta_0\|_2^2\leq s\mu^2$. To see this, note that $\mathbf{P}_{\pi|s}(\|\beta^*\|_2^2\leq s\mu^2)=1$ by the definition of $\pi_{|s}$ and thus, writing Π for the projection on $\{x\in\mathbb{R}^p:\|x\|_2\leq\sqrt{s}\mu\}$, we have with probability one under $\pi_{|s}$

$$\|\widehat{\theta}-\beta^*\|_2^2 \geq \|\Pi(\widehat{\theta})-\Pi(\beta^*)\|_2^2 = \|\Pi(\widehat{\theta})-\beta^*\|_2^2$$

for any estimator $\widehat{\theta}$, where the inequality is from the non-expansivity of Π, cf. [40], §E.9.3. Hence it suffices to consider the infimum over all estimators $\{\widehat{\theta}:\|\widehat{\theta}\|_2^2\leq s\mu^2\}$ in (1.74). We now have

$$\begin{aligned}
&B(\pi) = \inf_{\widehat{\theta}} \mathbf{E}_\pi\mathbf{E}_{X,\beta^*}[\|\widehat{\theta}-\beta^*\|_2^2]\\
&\leq \mathbf{E}_\pi\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]\\
&= \mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]I(\beta^*\in B_0^+(s;p))\right] + \mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]I(\beta^*\notin B_0^+(s;p))\right]\\
&= \mathbf{E}_{\pi|s}\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]\right]\mathbf{P}_\pi(\beta^*\in B_0^+(s;p)) + \mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]I(\beta^*\notin B_0^+(s;p))\right]\\
&\leq \mathbf{E}_{\pi|s}\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]\right] + \mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]I(\beta^*\notin B_0^+(s;p))\right]\\
&\leq R(B_0^+(s;p);X) + \mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]I(\beta^*\notin B_0^+(s;p))\right],
\end{aligned} \tag{1.75}$$

where the last inequality is because of (1.67) and the definition of $\pi_{|s}$. We now bound the second term.

$$\begin{aligned}
&\mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0-\beta^*\|_2^2]I(\beta^*\notin B_0^+(s;p))\right]\\
&\leq 2\mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0\|_2^2+\|\beta^*\|_2^2]I(\beta^*\notin B_0^+(s;p))\right]\\
&\leq 2s\mu^2\mathbf{P}_\pi(\beta^*\notin B_0^+(s;p)) + 2\mathbf{E}_\pi\left[\|\beta^*\|_2^2 I(\beta^*\notin B_0^+(s;p))\right]\\
&= 2s\mu^2\mathbf{P}_\pi(\|\beta^*\|_0>s) + 2\mu^2\mathbf{E}_\pi\left[\|\beta^*\|_0 I(\beta^*\notin B_0^+(s;p))\right]
\end{aligned} \tag{1.76}$$

Observe that under π, $\|\beta^*\|_0$ follows a Binomial distribution with p trials and probability of success α. For what follows, we choose $\alpha = \alpha_{p,s,\varepsilon} = (1-\varepsilon)s/p$ for $\varepsilon \in (0,1)$. We then have

$$\mathbf{E}_\pi[\|\beta^*\|_0] = (1-\varepsilon)s, \qquad \mathbf{Var}_\pi[\|\beta^*\|_0] \leq s.$$

From Bernstein's inequality ([106], Eq. (2.16)), we hence obtain that

$$\mathbf{P}_\pi(\|\beta^*\|_0 > s(1-\varepsilon)+t) \leq \begin{cases} \exp\left(-\frac{t^2}{4s}\right) & \text{if } t \leq 3s \\ \exp\left(-\frac{3}{4}t\right) & \text{otherwise.} \end{cases}$$

With the choice $t = s\varepsilon$, we obtain

$$\mathbf{P}_\pi(\|\beta^*\|_0 > s) \leq \exp\left(-\frac{s\varepsilon^2}{4}\right). \tag{1.77}$$

Let $j^\star = \max\{j : j - s(1-\varepsilon) \leq 3s\}$. We then also have

$$\begin{aligned}
\mathbf{E}_\pi\left[\|\beta^*\|_0 I(\beta^* \notin B_0^+(s;p))\right] &= \sum_{j=s+1}^{p} j\,\mathbf{P}_\pi(\|\beta^*\|_0 = j) \\
&= \sum_{j=s+1}^{j^\star} j\,\mathbf{P}_\pi(\|\beta^*\|_0 = j) + \sum_{j=j^\star+1}^{p} j\,\mathbf{P}_\pi(\|\beta^*\|_0 = j) \\
&\leq 4s\,\mathbf{P}_\pi(\|\beta^*\|_0 > s) + (j^\star+1)\sum_{j=j^\star+1}^{p} \mathbf{P}_\pi(\|\beta^*\|_0 \geq j) \\
&\leq 4s\exp\left(-\frac{s\varepsilon^2}{4}\right) + (4s+1)\sum_{j=j^\star+1}^{p} \exp\left(-\frac{3}{4}(j-s(1-\varepsilon))\right) \\
&\leq 4s\exp\left(-\frac{s\varepsilon^2}{4}\right) + (4s+1)\int_{3s}^{\infty} \exp\left(-\frac{3}{4}u\right)\,du \\
&= 4s\exp\left(-\frac{s\varepsilon^2}{4}\right) + (4s+1)\,\frac{4}{3}\exp\left(-\frac{9}{4}s\right)
\end{aligned} \tag{1.78}$$

Putting together the pieces. Inserting (1.77) and (1.78) into (1.76), we obtain

$$\mathbf{E}_\pi\left[\mathbf{E}_{X,\beta^*}[\|\delta_0 - \beta^*\|_2^2] I(\beta^* \notin B_0^+(s;p))\right] \leq s\mu^2 10\exp\left(-\frac{s\varepsilon^2}{4}\right) + (4s+1)\mu^2\frac{8}{3}\exp\left(-\frac{9}{4}s\right).$$

Plugging this bound into (1.75) yields

$$R(B_0^+(s;p);X) \geq B(\pi) - s\mu^2 10\exp\left(-\frac{s\varepsilon^2}{4}\right) - (4s+1)\mu^2\frac{8}{3}\exp\left(-\frac{9}{4}s\right).$$

We now let $\varepsilon = s^{-\kappa}$ for $0 < \kappa < 1/2$ and let μ satisfy (1.72) with $\alpha = \alpha_{p,s,\varepsilon} = s/p(1-\varepsilon)$ as above. Invoking (1.73) as deduced from Theorem 1.27, we have as

$$p/s \longrightarrow \infty, \qquad s \longrightarrow \infty$$

$$R(B_0^+(s;p);X) \geq (1+o(1))2\sigma^2\frac{s}{n}\log(p/s),$$

where we have used that

$$s\mu^2 10\exp\left(-\frac{s\varepsilon^2}{4}\right)+(4s+1)\mu^2\frac{8}{3}\exp\left(-\frac{9}{4}s\right) = o(1)2\sigma^2\frac{s}{n}\log(p/s) \quad \text{as } p/s\to\infty,\ s\to\infty.$$

This completes the proof of the theorem.

1.4.4 Estimation error with respect to the ℓ_∞-norm and support recovery by thresholding

In the present subsection, we directly derive bounds on the ℓ_∞-estimation error of NNLS using a different set of conditions as in §1.4.2. In light of these bounds, we subsequently study the performance of a thresholded NNLS estimator with regard to support recovery.

Separating hyperplane constant. The approach we pursue in the sequel builds on the geometry of $\mathcal{C}_X$ according to the discussion in §1.3.2 leading to Proposition 1.9, which implies that $\beta^* \in B_0^+(S;p)$ can be exactly recovered from $y = X\beta^*$ via NNLS if $\mathcal{C}_{X_S}$ is a face of $\mathcal{C}_X$. That is, there exists $w\in\mathbb{R}^n$ such that $X_S^\top w = 0$ and $X_{S^c}^\top w \succ 0$, where w can be interpreted as the normal vector of a separating hyperplane for the sets $\mathcal{C}_{X_S}$ and $\mathcal{C}_X\setminus\mathcal{C}_{X_S}$. The intuition is that in the presence of noise, in order to be able to expect that $\widehat{\beta}$ is close to β^*, the separation should be significant enough. This is quantified terms of the *separating hyperplane constant*

$$\tau(S) = \left\{\max\tau:\ \exists w\in\mathbb{R}^n,\ \|w\|_2\leq 1 \quad \text{s.t.}\ \ \frac{1}{\sqrt{n}}X_S^\top w = 0 \ \text{ and } \ \frac{1}{\sqrt{n}}X_{S^c}^\top w \succeq \tau\mathbf{1}\right\}. \tag{1.79}$$

Note that the definition of $\tau(S)$ parallels the definition of the constant τ_0 (1.46) used to define the self-regularizing property. 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. Defining

$$Z = \Pi_S^\perp X_{S^c}, \tag{1.80}$$

we have by convex duality

$$\begin{aligned}\tau^2(S) &= \min_{\substack{\theta\in\mathbb{R}^s\\ \lambda\in T^{p-s-1}}}\frac{1}{n}\left\|X_S\theta - X_{S^c}\lambda\right\|_2^2, \ \text{ where } T^{p-s-1} = \{\lambda\in\mathbb{R}_+^{p-s}:\ \lambda^\top\mathbf{1}=1\}\\ &= \min_{\lambda\in T^{p-s-1}}\lambda^\top\frac{1}{n}X_{S^c}^\top\Pi_S^\perp X_{S^c}\lambda = \min_{\lambda\in T^{p-s-1}}\lambda^\top\frac{1}{n}Z^\top Z\lambda.\end{aligned} \tag{1.81}$$

The last line highlights the connection to (1.47) in §1.4.1. Expanding $\frac{1}{n}Z^\top Z$ under the assumption that the submatrix Σ_{SS} is invertible, $\tau^2(S)$ can be written as

$$\tau^2(S) = \min_{\lambda\in T^{p-s-1}}\lambda^\top\left(\Sigma_{S^cS^c} - \Sigma_{S^cS}\left(\Sigma_{SS}\right)^{-1}\Sigma_{SS^c}\right)\lambda \tag{1.82}$$

Bounds on the ℓ_∞-error. The next theorem also covers the case of an approximately sparse target, and for once, S is here used to denote the set of the s largest coefficients of β^* instead of its support (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. To state the theorem, we need the quantities below, which also appear in the upper bound on the ℓ_∞-error of the lasso [163].

$$\beta_{\min}(S) = \min_{j\in S} \beta^*_j, \quad K(S) = \max_{\|v\|_\infty=1} \|(\Sigma_{SS})^{-1}v\|_\infty, \quad \phi_{\min}(S) = \min_{\|v\|_2=1} \|\Sigma_{SS}v\|_2\,. \tag{1.83}$$

We assume throughout that $\phi_{\min}(S) > 0$, or equivalently, that $(\Sigma_{SS})^{-1}$ exists; otherwise, estimation of β^*_S would be hopeless.

Theorem 1.28. *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*

$$b = \frac{2\left(\|\beta^*_{S^c}\|_1 + (1+M)\sigma\sqrt{\frac{2\log p}{n}}\right)}{\tau^2(S)}, \qquad \text{and } \widetilde{b} = (b + \|\beta^*_{S^c}\|_1)\cdot K(S) + \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(S)}}\sqrt{\frac{2\log p}{n}}. \tag{1.84}$$

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

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

A proof of Theorem 1.28 is provided in a separate paragraph below.
Theorem 1.28 can be summarized as follows; for convenience, it is assumed that $\beta^*_{S^c} = 0$. 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 §1.4.6, 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 1.24 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 1.24. On the positive side, Theorem 1.28 may yield a satisfactory result for s constant or growing only slowly with n, without requiring the restricted eigenvalue condition of Theorem 1.24.

Towards a possible improvement of Theorem 1.28. The potentially sub-optimal dependence on the sparsity level s in the bounds of Theorem 1.28 is too pessimistic relative to the empirical behaviour of NNLS as discussed in §1.4.8. The performance reported there can be better understood in light of Theorem 1.29 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 1.30 below. 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 1.29. *Let the data-generating model be as in Theorem 1.28, and assume additionally that $\beta^*_{S^c} = 0$. Let $M \geq 0$ be arbitrary. If the columns of X are in general linear position (1.17) and if*

$$\frac{32(1+M)^2\sigma^2}{\mathbf{E}[\varepsilon_1^2]} \; \frac{\log p}{\tau^2(S)\, n} \leq \left(1 - \frac{s}{n}\right), \tag{1.85}$$

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 (1.84), then $S \subseteq F$ and

$$\|\widehat{\beta} - \beta^*\|_\infty \leq \frac{(1+M)\sigma}{\sqrt{\phi_{\min}(F)}} \sqrt{\frac{2\log p}{n}}, \tag{1.86}$$

with probability at least $1 - 6p^{-M^2}$, where $\phi_{\min}(F)$ is defined analogously to $\phi_{\min}(S)$ in (1.83).

A proof of Theorem 1.29 is provided in a separate paragraph below.
We first note that for s/n bounded away from 1, condition (1.85) 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 1.28, 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 1.29 may yield a significantly improved bound on $\|\widehat{\beta} - \beta^*\|_\infty$ as compared to Theorem 1.28 if $\{\phi_{\min}(F)\}^{-1/2}$, the inverse of the smallest singular value of $X_F/\sqrt{n} \in \mathbb{R}^{n \times |F|}$, scales more favourably than $K(S)/\tau^2(S)$. In this context, note 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 [32, 110, 173]. That assumption is supported by results in random matrix theory [99, 133]. In §1.4.6 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 1.37 in §1.4.6).

Proof of Theorem 1.28. In order to prove Theorem, we state and prove two additional lemmas.

Lemma 1.30. *$\widehat{\beta}$ is a minimizer of the NNLS problem (1.16) if and only if there exists $F \subseteq \{1, \ldots, p\}$ such that*

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

Proof. The Lagrangian (cf. [16], §5.1.1) of (1.16) is given by

$$L(\beta, \mu) = \frac{1}{n}\|y - X\beta\|_2^2 - \mu^\top \beta, \tag{1.87}$$

where $\mu \succeq 0$ is a vector of Lagrangian multipliers. According to the Karush-Kuhn-Tucker (KKT) conditions (cf. [16], §5.5.3) associated with (1.87), $\widehat{\beta} \succeq 0$ is a minimizer of the NNLS problem if and only if there exists a Lagrangian multiplier $\widehat{\mu} \succeq 0$ such that

$$\frac{1}{n}X_j^\top(y - X\widehat{\beta}) = -\widehat{\mu}_j, \tag{1.88}$$

$$\widehat{\mu}_j\widehat{\beta}_j = 0, \quad j = 1, \ldots, p. \tag{1.89}$$

Let $F = \{j : \widehat{\beta}_j > 0\}$. In virtue of (1.89), for $j \in F$, it must hold that $\widehat{\mu}_j = 0$ and thus $\frac{1}{n}X_j^\top(y - X\widehat{\beta}) = 0$ according to (1.88). For $j \in F^c$, the assertion of the Lemma follows from (1.88) and the fact that $\widehat{\mu} \succeq 0$. □

Lemma 1.30 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

$$\frac{1}{n}\|y - X\widehat{\beta}\|_2^2 = \min_{\beta} \frac{1}{n}\|y - X\beta\|_2^2 \quad \text{subject to } \beta_{F^c} = 0. \tag{1.90}$$

Lemma 1.31. *In the situation of Theorem 1.28, let $Z = \Pi_S^\perp X_{S^c}$ as defined in (1.80) and $\xi = \Pi_S^\perp \varepsilon$. Consider the two non-negative least squares problems*

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

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

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 (1.16).

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

$$\begin{aligned}\frac{1}{n}\|y - X\beta\|_2^2 &= \frac{1}{n}\|\Pi_S y - \Pi_S X\beta\|_2^2 + \frac{1}{n}\left\|\Pi_S^\perp y - \Pi_S^\perp X\beta\right\|_2^2 \\ &= \frac{1}{n}\|\Pi_S\varepsilon + \Pi_S X_{S^c}\beta^*_{S^c} + X_S\beta^*_S - X_S\beta_S - \Pi_S X_{S^c}\beta_{S^c}\|_2^2 + \qquad (1.91) \\ &\quad + \frac{1}{n}\|\xi + Z\beta^*_{S^c} - Z\beta_{S^c}\|_2^2 \qquad (1.92)\end{aligned}$$

Separate minimization of the second summand (1.92) yields $\widehat{\beta}^{(P1)}$. Substituting $\widehat{\beta}^{(P1)}$ for β_{S^c} in the first summand (1.91), and minimizing the latter amounts to solving $(P2)$. In view of Lemma 1.30, 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 $X_S\beta^*_S + \Pi_S(\varepsilon + X_{S^c}\beta^*_{S^c} - X_{S^c}\widehat{\beta}^{(P1)})$ of the non-negative least squares problem $(P2)$ is contained in the column space of X_S. Since the second summand (1.92) corresponding to $(P1)$ cannot be made smaller than by separate minimization, we have minimized the non-negative least squares objective. □

Proof. (Theorem 1.28) Consider problem $(P1)$ of Lemma 1.31.
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

$$\begin{aligned}
(\widehat{\beta}^{(P1)})^\top \frac{1}{n} Z^\top Z \widehat{\beta}^{(P1)} &\leq \|\widehat{\beta}^{(P1)}\|_1 \left(A + 2 \left\| \tfrac{1}{n} Z^\top Z \beta^*_{S^c} \right\|_\infty \right), \quad A := \max_j \frac{2}{n} |Z_j^\top \xi|. \\
&\leq \|\widehat{\beta}^{(P1)}\|_1 \left(A + 2\|\beta^*_{S^c}\|_1 \max_{j,k} Z_j^\top Z_k / n \right) \\
&\leq \|\widehat{\beta}^{(P1)}\|_1 \left(A + 2 \max_{j,k} \|Z_j/\sqrt{n}\|_2 \|Z_k/\sqrt{n}\|_2 \|\beta^*_{S^c}\|_1 \right) \\
&\leq \|\widehat{\beta}^{(P1)}\|_1 (A + 2\|\beta^*_{S^c}\|_1).
\end{aligned} \tag{1.93}$$

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

$$\|Z_j\|_2 = \|\Pi_S^\perp X_j\|_2 \leq \|X_j\|_2. \tag{1.94}$$

As observed in (1.81), $\tau^2(S) = \min_{\lambda \in T^{p-s-1}} \lambda^\top \frac{1}{n} Z^\top Z \lambda$, s.t. the l.h.s. of (1.93) can be lower bounded via

$$(\widehat{\beta}^{(P1)})^\top \frac{1}{n} Z^\top Z \widehat{\beta}^{(P1)} \geq \tau^2(S) \|\widehat{\beta}^{(P1)}\|_1^2. \tag{1.95}$$

Combining (1.93) and (1.95), 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 1.31, 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 1.31 then yields $\bar{\beta}^{(P2)} = \widehat{\beta}^{(P2)} = \widehat{\beta}_S$. Using the closed form expression for the ordinary least squares estimator, one obtains

$$\begin{aligned}
\bar{\beta}^{(P2)} &= \frac{1}{n} (\Sigma_{SS})^{-1} X_S^\top \Pi_S (y - X_{S^c} \widehat{\beta}^{(P1)}) \\
&= \frac{1}{n} (\Sigma_{SS})^{-1} X_S^\top (X_S \beta^*_S + \Pi_S \varepsilon - \Pi_S X_{S^c} (\widehat{\beta}^{(P1)} - \beta^*_{S^c})) \\
&= \beta^*_S + \frac{1}{n} (\Sigma_{SS})^{-1} X_S^\top \varepsilon - (\Sigma_{SS})^{-1} \Sigma_{SS^c} (\widehat{\beta}^{(P1)} - \beta^*_{S^c}).
\end{aligned} \tag{1.96}$$

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

$$\begin{aligned}\|(\Sigma_{SS})^{-1}\Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})\|_\infty &\leq \max_{\|v\|_\infty=1}\|(\Sigma_{SS})^{-1}v\|_\infty\|\Sigma_{SS^c}(\widehat{\beta}^{(P1)} - \beta^*_{S^c})\|_\infty \\ &\overset{(1.83)}{\leq} K(S)\,(\|\widehat{\beta}^{(P1)}\|_1 + \|\beta^*_{S^c}\|_1).\end{aligned} \tag{1.97}$$

Step 3: *Putting together the pieces.* The two random terms A and A_S are maxima of a finite collection of linear combinations of zero-mean sub-Gaussian random variables so that (1.19) can be applied by estimating Euclidean norms. For A, we use (1.94). Second, we have

$$A_S = \max_{1\leq j\leq s}\frac{|v_j^\top\varepsilon|}{n}, \quad v_j = X_S(\Sigma_{SS})^{-1}e_j, \quad j = 1,\dots,s, \tag{1.98}$$

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

$$\max_{1\leq j\leq s}\|v_j\|_2^2 = \max_{1\leq j\leq s} e_j^\top(\Sigma_{SS})^{-1}X_S^\top X_S(\Sigma_{SS})^{-1}e_j \overset{(1.83)}{\leq} \frac{n}{\phi_{\min}(S)}.$$

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 1.28, 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 $\widehat{\beta}^{(P1)} = \widehat{\beta}_{S^c}$. □

Proof of Theorem 1.29. 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 [89].

Lemma 1.32. *[89] Let $\Pi \in \mathbb{R}^{n\times n}$ be a projection matrix on a d-dimensional subspace of $\mathbb{R}^n$ and let $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n)^\top$ 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 1.33. *Let the columns of X be in general linear position. Then the NNLS problem has a unique solution if one of the following holds:*

$$\textit{(i)}\ \ p \le n, \qquad \textit{(ii)}\ \ \min_{\beta \succeq 0} \tfrac{1}{n}\|y - X\beta\|_2^2 > 0.$$

Moreover, under (ii) the active set $F = \{j : \widehat{\beta}_j > 0\}$ satisfies $|F| \le \min\{n-1, p\}$. Conversely, if y has a distribution that is absolutely continuous with respect to the Lebesgue measure, then $|F| \le \min\{n-1, p\}$ 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 equals the projection of y onto $\mathcal{C}_X$, which is a closed convex set (cf. [40], appendix E.9). Moreover, under (ii), $X\widehat{\beta}$ must be contained in $\mathbf{bd}\,\mathcal{C}_X$ (by general linear position, $\mathbf{int}\,\mathcal{C}_X$ is non-empty). At this point, we resort to the characterization of $\mathbf{bd}\,\mathcal{C}_X$ in §1.3.2. Accordingly, $\mathbf{bd}\,\mathcal{C}_X$ equals the union of the facets of $\mathcal{C}_X$. Under general linear position, each of the facets is given by a polyhedral cone $\mathcal{C}_{X_J}$ for $J \in \mathcal{J}(n-1)$, and all points contained in $\mathbf{bd}\,\mathcal{C}_X$ have a unique representation as non-negative combination of $\{X_j\}_{j=1}^p$.
For the last part of the lemma, 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). □

Proof. (Theorem 1.29)
Part 1: proof of uniqueness. Using Lemma 1.33 and condition (1.85), which reads

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

we will show that for $p \ge n$, condition (ii) of Lemma 1.33 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 1.31 specialized to the case $\beta^*_{S^c} = 0$. 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

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

Hence, conditional on the event

$$\left\{\|\Pi_S^{\perp}\varepsilon\|_2^2 > \mathbf{E}[\varepsilon_1^2]\,\frac{n-s}{4}\right\} \cap \left\{\left(\frac{2}{n}\|Z^{\top}\xi\|_{\infty}\right)^2 \leq 8(1+M)^2\sigma^2\,\frac{\log p}{n}\right\} \tag{1.99}$$

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 (1.85). As a result, $\min_{\beta\succeq 0}\frac{1}{n}\|y-X\beta\|_2^2 > 0$ as was to be shown. Invoking Lemma 1.32 with $\Pi = \Pi_S^{\perp}$ so that $d = n-s$ by general linear position and treating the second event in (1.99) as in step 3 of the proof of Theorem 1.28, the probability of the event (1.99) 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 1.30) 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 (1.98) to conclude that for any $M \geq 0$, the event

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

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 used for the proof of Theorem 1.28 . Given the lower bound on $\beta_{\min}(S)$, it is established therein 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. □

Support recovery by thresholding. The bounds on the estimation error presented above 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

$$\widehat{\beta}_j(t) = \begin{cases}\widehat{\beta}_j, & \text{if } \widehat{\beta}_j > t,\\ 0, & \text{otherwise, } j = 1,\ldots,p,\end{cases} \tag{1.100}$$

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 1.28 or 1.29: 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 [65] for support recovery on the basis of marginal regression. A central observation in [65] 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 1.24 to 1.29, NNLS may give rise to such ranking by setting

$$r_j = k \iff \widehat{\beta}_j = \widehat{\beta}_{(k)},\ j = 1, \ldots, p, \tag{1.101}$$

where $\widehat{\beta}_{(1)} \geq \widehat{\beta}_{(2)} \geq \ldots \geq \widehat{\beta}_{(p)}$ is the sequence of coefficients arranged in decreasing order. Theorem 1.34 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 [65]. Unlike the corresponding result in [65], our statement is non-asymptotic and comes with a more transparent condition on the design and $\beta_{\min}(S)$. We point out that Theorem 1.34 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 1.34. *Consider the data-generating model of Theorem 1.29 and suppose that the NNLS estimator has the property that according to (1.101), it holds that $r_j \leq s$ for all $j \in S$. For any $M \geq 0$, set*

$$\begin{aligned} \widehat{s} &= \max\left\{0 \leq k \leq (p-1) : \delta(k) \geq (1+M)\sigma\sqrt{2\ \log p}\right\} + 1, \\ \text{where}\ \ \delta(k) &= \|(\Pi(k+1) - \Pi(k))y\|_2\,,\ k = 0, \ldots, (p-1), \end{aligned} \tag{1.102}$$

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 $1 - 4p^{-M^2}$.

Proof. We first recall that the analysis is conditional on the event

$$E = \{r_j \leq s \text{ for all } j \in S\}, \quad \text{where } r_j = k \Leftrightarrow \widehat{\beta}_j = \widehat{\beta}_{(k)}. \tag{1.103}$$

Our proof closely follows the corresponding proof in [65]. 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

$$\|(\Pi(k+1) - \Pi(k))\varepsilon\|_2 = |\,\langle u_k, \varepsilon\rangle\,|, \quad k = 0, \ldots, p-1. \tag{1.104}$$

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

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

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 $\{S \backslash \widehat{S} = \emptyset\}$ is contained in the event $\{\delta(s-1) \ge \Delta\}$. Hence it suffices to upper bound the probability of the event $\{\delta(s-1) < \Delta\}$. We have

$$\begin{aligned}
\mathbf{P}(\delta(s-1) < \Delta) &= \mathbf{P}\left(\|(\Pi(s) - \Pi(s-1))y\|_2 < \Delta\right) \\
&\le \mathbf{P}\left(\|(\Pi(s) - \Pi(s-1))X_S\beta_S^*\|_2 < \Delta + \|(\Pi(s) - \Pi(s-1))\varepsilon\|_2\right) \\
&\overset{(1.104)}{=} \mathbf{P}\left(\|(\Pi(s) - \Pi(s-1))X_S\beta_S^*\|_2 < \Delta + |\langle u_{s-1}, \varepsilon\rangle|\right) \\
&\le \mathbf{P}\left(\min_{j\in S}\|(\Pi_S - \Pi_{S\backslash j})X_j\beta_j^*\|_2 < \Delta + |\langle u_{s-1}, \varepsilon\rangle|\right), \qquad (1.105)
\end{aligned}$$

where Π_S and $\Pi_{S\backslash j}$ 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

$$\min_{j\in S}\|(\Pi_S - \Pi_{S\backslash j})X_j\beta_j^*\|_2 \ge \sqrt{n}\left\{\phi_{\min}(S)\right\}^{1/2}\beta_{\min}(S). \qquad (1.106)$$

Combining (1.105) and (1.106), it suffices to upper bound

$$\mathbf{P}\left(|\langle u_{s-1}, \varepsilon\rangle| > \sqrt{n}\left\{\phi_{\min}(S)\right\}^{1/2}\beta_{\min}(S) - \Delta\right) \qquad (1.107)$$

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

$$\begin{aligned}
\mathbf{P}(\cup_{k=s}^{p-1}\{\delta(k) \ge \Delta\}) &= \mathbf{P}\left(\max_{s\le k\le p-1}\|(\Pi(k+1) - \Pi(k))y\|_2 \ge \Delta\right) \\
&= \mathbf{P}\left(\max_{s\le k\le p-1}\|(\Pi(k+1) - \Pi(k))\varepsilon\|_2 \ge \Delta\right) \\
&= \mathbf{P}\left(\max_{s\le k\le p-1}|\langle u_k, \varepsilon\rangle| \ge \Delta\right). \qquad (1.108)
\end{aligned}$$

Choosing $\Delta = (1+M)\sigma\sqrt{2\log(p)}$ for an arbitrary $M \ge 0$, using the assumption on $\beta_{\min}(S)$, and controlling $\max_{0\le k\le p-1}|\langle u_k, \varepsilon\rangle|$ according to (1.19) in the usual way, the probabilities (1.107) and (1.108) do not exceed $2p^{-M^2}$. The assertion of the theorem then follows. To conclude the proof, it remains to establish (1.106). Let us fix an arbitrary $j \in S$. We have

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

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

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

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

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

□

We note that the lower bound on $\beta_{\min}(S)$ required in Theorem 1.34 is rather moderate. Similar or more stringent lower bounds are required throughout the literature on support recovery in a noisy setup [21, 28, 100, 163, 175, 176], and are typically already needed to ensure that the variables in S are ranked at the top (cf. also Theorems 1.24 to 1.29).
Strictly speaking, the estimator $\widehat{s}$ in Theorem 1.34 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 [78]. 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 [8](cf. §1.4.8).
A nice property of the approach is its computational simplicity. Repeated evaluation of $\delta(k)$ in (1.102) 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.

1.4.5 Comparison with the non-negative lasso

Let us recall the non-negative lasso (1.15)

$$\widehat{\beta}^{\ell_1^+,\lambda} \in \operatorname*{argmin}_{\beta\in\mathbb{R}_+^p} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda \mathbf{1}^\top\beta,\ \ \lambda > 0.$$

In the present subsection, we elaborate on the performance of the non-negative lasso with regard to estimation in ℓ_∞-norm and support recovery. The analysis is in correspondence with that of the unconstrained lasso in [163] and reveals that the non-negativity constraints do not lead to qualitative differences.
While support recovery can be achieved in a one-stage manner (i.e. without subsequent thresholding), it requires a restrictive condition on the design. This deficiency persists

[8]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).

irrespectively of the scaling of n, p, s and $\beta_{\min}(S)$. In this regard, ℓ_1-regularization behaves rather differently from ℓ_0-regularization, which achieves support recovery under a minimal set of conditions (cf. Theorem 4 in [174] and Proposition 1.1). A second shortcoming of the non-negative lasso is that in general, it does not attain the optimal rate in estimation with respect to the ℓ_∞-norm. These negative findings do not cast doubts about the usefulness of the (non-negative) lasso for high-dimensional sparse regression in general, but they indicate that there is room left for improvements. This provides motivation to consider alternative methods.

Non-negative irrepresentable condition. For given support S, the *non-negative irrepresentable constant* is defined as

$$\iota(S) = \max_{j \in S^c} \Sigma_{jS}(\Sigma_{SS})^{-1}\mathbf{1} = \max_{j \in S^c} X_j^\top X_S (X_S^\top X_S)^{-1}\mathbf{1}. \tag{1.109}$$

As stated in the next proposition, the non-negative irrepresentable condition $\iota(S) < 1$ is necessary for the non-negative lasso to recover the support S with a significant level of confidence.

Proposition 1.35. *Assume that $y = X\beta^* + \varepsilon$, where $\beta^* \succeq 0$ has support S and ε has independent entries with zero mean. If the non-negative irrepresentable constant (1.109) satisfies $\iota(S) \geq 1$, then for any $\lambda > 0$, the non-negative lasso obeys*

$$\mathbf{P}\left(\{j : \widehat{\beta}_j^{\ell_1^+,\lambda} > 0\} = S\right) \leq 1/2.$$

Proposition 1.35 is proved collectively with Proposition 1.36 below. The negative result of the former may appear surprising, because it holds irrespectively of $\beta_{\min}(S)$. Eventually, this issue is a consequence of the geometry of the non-negative lasso problem. It can be formulated equivalently as the projection problem

$$\min_z \|y - z\|_2^2 \quad \text{subject to } z \in \gamma(\lambda)\mathcal{P}_{0,X}, \tag{1.110}$$

where $\mathcal{P}_{0,X}$ denotes the convex hull of $\{0\} \cup \{X_j\}_{j=1}^p$ as introduced in (1.32) and $\gamma(\lambda)$ is a bound on $\mathbf{1}^\top \widehat{\beta}^{\ell_1^+,\lambda}$ that depends on the regularization parameter λ. In terms of (1.110), the requirement of having no false positive selections essentially asks for the following: we are given a point $X_S\beta_S^*$ contained in a face of $(\mathbf{1}^\top\beta_S^*)\,\mathcal{P}_{0,X}$ and perturb it by adding ε and compute the projection on $\gamma(\lambda)\mathcal{P}_{0,X}$, which then must be contained in $\gamma(\lambda)\mathcal{P}_{0,X_S}$. The requirement that $X_S\beta_S^*$ be contained in a face of $(\mathbf{1}^\top\beta_S^*)\,\mathcal{P}_{0,X}$ is obviously necessary in this context, because otherwise support recovery already fails even without noise (cf. §1.3.3). Accordingly, the irrepresentable condition $\iota(S) < 1$ implies that $\mathcal{P}_{X_S}$ is a face of $\mathcal{P}_{0,X}$. Indeed, setting $w = -X_S(X_S^\top X_S)^{-1}\mathbf{1}$ and $b = -1$, we find that the condition given in (1.33) is satisfied for $J = S$. Conversely, it is easy to find examples where $\mathcal{P}_{X_S}$ is a face of $\mathcal{P}_{0,X}$, but the irrepresentable condition fails. We here present an example from [180]. Consider $X = [X_1\, X_2\, X_3]$ such that the Gram matrix is given by

$$\Sigma = \begin{pmatrix} 1 & 0 & 2/3 \\ 0 & 1 & 2/3 \\ 2/3 & 2/3 & 1 \end{pmatrix}$$

Let $S = \{1, 2\}$. Then it is easy to see that $\iota(S) = 4/3 > 1$. On the other hand, Σ is strictly positive definite, which implies that $\{X_1, X_2, X_3\}$ are linearly and $\{0, X_1, X_2, X_3\}$ affinely independent, respectively. As a result, $\mathcal{P}_{0,X}$ is simplicial, and in particular, $\mathcal{P}_{X_S}$ is a face of $\mathcal{P}_{0,X}$.

Upper bound on the ℓ_∞-error in estimation and support recovery. Under the non-negative irrepresentable condition, a suitable choice of the regularization parameter and an according lower bound on $\beta_{\min}(S)$, one can prove that the non-negative lasso achieves support recovery. As a by-product, one also obtains an upper bound on $\|\widehat{\beta}^{\ell_1^+,\lambda} - \beta^*\|_\infty$.

Proposition 1.36. *Assume that $y = X\beta^* + \varepsilon$, where $\beta^* \succeq 0$ has support S and ε has i.i.d. zero-mean sub-Gaussian entries with parameter σ. Suppose further that the non-negative irrepresentable condition $\iota(S) < 1$ according to (1.109) holds. For any $M \geq 0$, if*

$$\begin{aligned} &\lambda > \frac{2\lambda_M}{1 - \iota(S)}, \quad \text{where} \quad \lambda_M = (1 + M)\,\sigma\sqrt{\frac{2\log p}{n}}, \\ \text{and} \quad &\beta_{\min}(S) > b, \quad \text{where} \quad b = \frac{\lambda}{2}\|(\Sigma_{SS})^{-1}\mathbf{1}\|_\infty + \frac{\lambda_M}{\sqrt{\phi_{\min}(S)}}, \end{aligned} \tag{1.111}$$

then $\{j : \widehat{\beta}_j^{\ell_1^+,\lambda} > 0\} = S$ and $\|\widehat{\beta}_S^{\ell_1^+,\lambda} - \beta_S^*\|_\infty \leq b$ with probability at least $1 - 4p^{-M^2}$.

Proof. (Proposition 1.35 and 1.36) Arguing similarly as in the proof of Lemma 1.30, we obtain that $\widehat{\beta}^{\ell_1^+,\lambda} \in \operatorname{argmin}_{\beta \in \mathbb{R}^p_+} n^{-1}\|y - X\beta\|_2^2 + \lambda\mathbf{1}^\top\beta$ if and only if there exists $\widehat{\mu} \succeq 0$ such that

$$\begin{aligned} &\frac{2}{n}X^\top(X\widehat{\beta}^{\ell_1^+,\lambda} - y) + \lambda\mathbf{1} = \widehat{\mu}, \\ &\widehat{\mu}_j\widehat{\beta}_j^{\ell_1^+,\lambda} = 0, \ j = 1, \ldots, p. \end{aligned} \tag{1.112}$$

The remainder of the proof is an adaptation of the scheme used in [163] to prove a corresponding result for the unconstrained lasso. Consider the following constrained non-negative lasso problem

$$\min_{\beta_S \succeq 0,\, \beta_{S^c} = 0} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\mathbf{1}^\top\beta, \tag{1.113}$$

which has a unique minimizer, say $\widehat{\alpha}$, once $\phi_{\min}(S) > 0$. Note that support recovery is achieved if and only if $\widehat{\beta}^{\ell_1^+,\lambda} = \widehat{\alpha}$ and $\widehat{\alpha}_S \succ 0$. Hence, in order to prove Proposition 1.35, it suffices to work conditional on the event $\{\widehat{\alpha}_S \succ 0\}$, plug in $\widehat{\alpha}$ for $\widehat{\beta}^{\ell_1^+,\lambda}$ in (1.112), and upper bound the (conditional) probability that $\widehat{\nu}_{S^c} \succeq 0$, where

$$\widehat{\nu} = \frac{2}{n}X^\top(X\widehat{\alpha} - y) + \lambda\mathbf{1}. \tag{1.114}$$

In fact, once there exists $j \in S^c$ such that $\widehat{\nu}_j < 0$, we cannot have $\widehat{\beta}^{\ell_1^+,\lambda} = \widehat{\alpha}$ in light of the optimality conditions (1.112) and thus do not achieve support recovery. Under

the assumptions made in the proposition, conditional on the event $\{\widehat{\alpha}_S \succ 0\}$, we have

$$\widehat{\alpha}_S = \beta^*_S - \frac{\lambda}{2}(\Sigma_{SS})^{-1}\mathbf{1} + (\Sigma_{SS})^{-1}\frac{1}{n}X_S^\top \varepsilon. \tag{1.115}$$

Substituting this expression back into (1.114), we obtain

$$\widehat{\nu}_{S^c} = -\lambda\left\{\Sigma_{S^cS}(\Sigma_{SS})^{-1}\mathbf{1}\right\} - \frac{2}{n}X_{S^c}^\top(I - \Pi_S)\varepsilon + \lambda\mathbf{1}. \tag{1.116}$$

Since $\iota(S) \geq 1$, there exists $j^* \in S^c$ such that $\Sigma_{j^*S}(\Sigma_{SS})^{-1}\mathbf{1} \geq 1$ and hence

$$\widehat{\nu}_{j^*} \leq -\frac{2}{n}X_{j^*}^\top(I - \Pi_S)\varepsilon.$$

Since the entries of ε are independent and have mean zero, the probability that $\widehat{\nu}_{j^*}$ is negative is at least one half, which concludes the proof of Proposition 1.35. For the proof of Proposition 1.36, we will argue in an opposite way. First, using that $\iota(S) < 1$, we will establish that $\widehat{\nu}_{S^c} \succ 0$ with high probability. Based on (1.116), we have

$$\min_{j\in S^c}\widehat{\nu}_j \geq \lambda(1 - \iota(S)) - \max_{j\in S^c}\frac{2}{n}X_j^\top(I - \Pi_S)\varepsilon > 2\lambda_M - \max_{j\in S^c}\frac{2}{n}X_j^\top(I - \Pi_S)\varepsilon.$$

In light of (1.19), the event

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

occurs with probability at least $1 - 2p^{-M^2}$, noting that $\|(I - \Pi_S)X_j\|_2 \leq \|X_j\|_2$ for all $j = 1, \ldots, p$. Second, we show that equipped with the lower bound on $\beta_{\min}(S)$, it holds that $\widehat{\alpha}_S \succ 0$ with probability at least $1 - 2p^{-M^2}$. Altogether, it then follows that $(\widehat{\alpha}, \widehat{\nu})$ satisfy the optimality conditions (1.112) of the non-negative lasso problem, which implies support recovery. In virtue of (1.115), we have

$$\|\widehat{\alpha}_S - \beta^*_S\|_\infty \leq \frac{\lambda}{2}\|(\Sigma_{SS})^{-1}\mathbf{1}\|_\infty + \left\|(\Sigma_{SS})^{-1}\tfrac{1}{n}X_S^\top\varepsilon\right\|_\infty.$$

Handling the random term on the r.h.s. as (1.98) in the proof of Theorem 1.28, we get that $\|\widehat{\alpha}_S - \beta^*_S\|_\infty \leq b$ with b as defined in Proposition 1.36. Since it is assumed that $\beta_{\min}(S) > b$, we finally obtain $0 \prec \widehat{\alpha}_S = \widehat{\beta}_S^{\ell_1^+,\lambda}$. □

There is some resemblance of the bound b in (1.111) and that of Theorem 1.28 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 (1.83). On the other hand, Proposition 1.36 yields a considerably stronger control of the off-support coefficients ($\widehat{\beta}_{S^c}^{\ell_1^+,\lambda} = 0$) as does Theorem 1.28, which only provides an ℓ_1-bound on $\widehat{\beta}_{S^c}$.

Suboptimality of the non-negative lasso with regard to estimation in ℓ_∞-norm and support recovery. Irrepresentable conditions as in Proposition 1.36 are regarded as rather restrictive in the literature [110, 173, 176]. Even in case the condition $\iota(S) < 1$ is fulfilled, the choice of λ in (1.111) with $\iota(S)$ possibly close to one may impose a rather stringent lower bound on $\beta_{\min}(S)$ so that support recovery can be achieved. At the same time, the choice $\lambda = 2\sigma\sqrt{2\log(p)/n}$ in combination with the restricted eigenvalue condition (Condition 1.23), which is regarded as far less restrictive than the irrepresentable condition, only yields a bound on $\|\widehat{\beta}^{\ell_1^+,\lambda} - \beta^*\|_q$ for $q \in [1,2]$ (cf. the discussion below Theorem 1.24), and it is no longer guaranteed that $\widehat{\beta}^{\ell_1^+,\lambda}_{S^c} = 0$. As a result, two-stage procedures like subsequent thresholding of $\widehat{\beta}^{\ell_1^+,\lambda}$ may be needed for support recovery. Moreover, support recovery (with or without subsequent thresholding) in general entails a sub-optimal condition on $\beta_{\min}(S) = \Omega(\sqrt{s\log(p)/n})$ because of the term $\|(\Sigma_{SS})^{-1}\mathbf{1}\|_\infty$ in (1.111) scaling as $\Theta(\sqrt{s})$ in the worst case. Let us consider a specific example in which the Gram matrix is of the form

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & 0 \\ 0 & I_{p-s} \end{bmatrix}, \quad \text{where } \Sigma_{SS} = \begin{bmatrix} 1 & -1/\sqrt{2(s-1)}\mathbf{1}_{s-1}^\top \\ -1/\sqrt{2(s-1)}\mathbf{1}_{s-1} & I_{s-1} \end{bmatrix}. \quad (1.117)$$

The constant $\sqrt{2}$ in the denominator is chosen for convenience; our argument would go through with any other constant larger than 1. Using Schur complements, one computes that

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

In particular, one has

$$e_1^\top(\Sigma_{SS})^{-1}\mathbf{1} = 2 + \sqrt{2(s-1)} = \Omega(\sqrt{s}), \quad (1.118)$$

Given the specific form of the Gram matrix in (1.117), one has that

$$\widehat{\beta}_j^{\ell_1^+,\lambda} = \max\{X_j^\top\varepsilon/n - \lambda/2, 0\}, \ \ j \in S^c.$$

To avoid false positive selections, λ has to be chosen such that $\max_{j\in S^c} 2X_j^\top\varepsilon/n < \lambda$. If ε has i.i.d. zero-mean Gaussian entries with variance σ^2, then $z = X_{S^c}^\top\varepsilon/n$ has i.i.d. zero-mean Gaussian entries with variance σ^2/n. Letting $p \to \infty$ while $s/p \le c < 1$, it holds that $\max_{1\le j\le p-s} z_j = \Theta(\sqrt{\log(p)/n})$ with probability tending to one (e.g. [80], §8.3). We deduce that the scaling $\lambda = \Omega(\sqrt{\log(p)/n})$ is required. This scaling of λ in turn entails the scaling $\beta_{\min}(S) = \Omega(\sqrt{s\log(p)/n})$ so that $\widehat{\beta}_S^{\ell_1^+,\lambda} \succ 0$ is ensured. Without loss of generality, suppose that $\beta_{\min}(S) = \beta_1^*$. Following the proof of Propositions 1.35/1.36 and using that $\Sigma_{SS^c} = 0$, we have

$$\begin{aligned} \widehat{\beta}_S^{\ell_1^+,\lambda} \succ 0 \iff & \beta_S^* - \frac{\lambda}{2}(\Sigma_{SS})^{-1}\mathbf{1} + (\Sigma_{SS})^{-1}\frac{1}{n}X_S^\top\varepsilon \succ 0 \\ \implies & \beta_{\min}(S) - \frac{\lambda}{2}e_1^\top(\Sigma_{SS})^{-1}\mathbf{1} + (\Sigma_{SS})^{-1}\frac{1}{n}X_S^\top\varepsilon > 0 \\ \implies & \beta_{\min}(S) > \frac{\lambda}{2}e_1^\top(\Sigma_{SS})^{-1}\mathbf{1} - \|(\Sigma_{SS})^{-1}X_S^\top\varepsilon/n\|_\infty. \end{aligned}$$

One computes that $\phi_{\min}(S) = 1 - 1/\sqrt{2}$, so that the random term on the right hand side scales as $O_{\mathbf{P}}(\sqrt{\log(p)/n})$. Consequently, using (1.118) and $\lambda = \Omega(\sqrt{\log(p)/n})$, we obtain the implication $\beta_{\min}(S) = \Omega(\sqrt{s\log(p)/n})$. Note that this scaling arises from the regularizer, and is thus not incurred by NNLS. In fact, setting $\lambda = 0$ in the first line of the above display, we obtain $\widehat{\beta}_S \succ 0$ once $\beta_{\min}(S) = \Omega(\sqrt{\log(p)/n})$ and support recovery can be achieved by thresholding at the same level.
While the example has been constructed with the intention to derive a suboptimal rate of the non-negative lasso with regard to estimation in the ℓ_∞-norm, some general insights are conveyed. The (non-negative) lasso is typically applied with a sufficiently large value of λ with the goal to have few non-zero entries in the solution $\widehat{\beta}^{\ell_1^+,\lambda}$. This is achieved at the cost of shrinking all entries towards zero, which may result into a substantial bias for the entries corresponding to S. Hence, in situations where such shrinkage is not indispensable, thresholding a NNLS solution may be a more favourable way to obtain a sparse model.

1.4.6 Discussion of the analysis of NNLS for selected designs

Our main results concerning the performance of NNLS as stated in Theorems 1.21 to 1.29 are subject to the following conditions: the self-regularizing property (Theorem 1.21), a combination of that property with a restricted eigenvalue condition (Theorem 1.24), a lower bound on the separating hyperplane constant (Theorem 1.28), and sparsity of the NNLS solution (Theorem 1.29). In the present section, we discuss to what extent these conditions are fulfilled for selected designs, some of which have already been considered in our analysis of the noiseless case (cf. §1.3.4). We here consider three basic classes. For the class of *non-self-regularizing designs* non-negativity constraints on the regression coefficients do not seem to yield any significant advantage. This is in contrast to the class of *equi-correlation-like designs*, which are shown to be tailored to NNLS. The third class comprises designs with a block or band structure arising in typical applications.

Non-self regularizing designs. In this paragraph, we provide several common examples of designs not having the self-regularizing property of Condition 1.19. Consequently, our main results, which rely on that condition, do not apply. Those designs can be identified by evaluating the quantity τ_0^2 (1.47) underlying Condition 1.19. From

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda \leq \frac{1}{p^2} \mathbf{1}^\top \Sigma \mathbf{1}, \tag{1.119}$$

we see that the sum of the entries of Σ must scale as $\Omega(p^2)$ for Condition 1.19 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: random matrices from the centrosymmetric ensemble
Recall the centrosymmetric ensemble (1.38) which comprises, roughly speaking, random matrices whose entries are drawn i.i.d. from a distribution symmetric around the

origin. Members of this class fail to satisfy condition $(\mathcal{H})$ with high probability once $p/n > 2$, in which case $\tau_0^2 = 0$.

Example 2: random matrices from $\texttt{HALFSPACE}(n, p, t)$ and ensemble $\mathrm{Ens}_{\mathbf{1}}(n, p)$
Recall from §1.3.4 that Construction $\texttt{HALFSPACE}(n, p, t)$ yields a design matrix composed of Gaussian random vectors contained in the interior of a halfspace $\{z \in \mathbb{R}^n : \langle z, w\rangle > 0\}$ for a given $w \in \mathbb{R}^n$. While condition $(\mathcal{H})$ is satisfied by construction, it can be shown that $\tau_0^2 = O_{\mathbf{P}}(1/n)$ for $p > 2n$; the same is true for instances of $\mathrm{Ens}_{\mathbf{1}}(n, p)$, cf. §1.3.4.

Example 3: orthonormal design
As already mentioned while motivating the self-regularizing property in §1.4.1, for $\Sigma = I$, τ_0^2 attains the upper bound in (1.119) which yields $\tau_0^2 = 1/p$.

Example 4: 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

$$\max_{1 \le j \le p} \sum_{k=1}^{p} \sigma_{jk} \le 2 \sum_{l=0}^{p-1} \rho^l \le 2(1-\rho)^{-1}$$

and (1.119) it follows that $\tau_0^2 \le 2p^{-1}(1-\rho)^{-1}$.

In all these examples, a similar reasoning applies with regard to the scaling of the separating hyperplane constant $\tau^2(S)$, regardless of the choice of S. In view of (1.81)

$$\tau^2(S) = \min_{\substack{\theta \in \mathbb{R}^s \\ \lambda \in T^{p-s-1}}} \frac{1}{n} \left\| X_S \theta - X_{S^c} \lambda \right\|_2^2 \le \min_{\lambda \in T^{p-s-1}} \frac{1}{n} \left\| X_{S^c} \lambda \right\|_2^2,$$

and the r.h.s. can be upper bounded via arguments used above. For example, this yields $\tau^2(S) \le 2(p-s)^{-1}(1-\rho)^{-1}$ (uniformly in S) when applied to power decay.

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. 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 (1.119) above.

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda \ge \min_{\lambda \in T^{p-1}} \lambda^\top \left\{ \sigma_0 \mathbf{1}\mathbf{1}^\top \right\} \lambda = \sigma_0, \tag{1.120}$$

i.e. Condition 1.19 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 1.19 is satisfied with $\tau^2 = \sigma_0/K$:

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda \ge \min_{\lambda \in T^{p-1}} \sum_{l=1}^{K} \lambda_{B_l}^\top \Sigma_{B_l B_l} \lambda_{B_l} \ge \sigma_0 \min_{\lambda \in T^{p-1}} \sum_{l=1}^{K} (\lambda_{B_l}^\top \mathbf{1})^2 = \sigma_0 / K, \tag{1.121}$$

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 1.2, the lower bound (1.121) 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 $X = (\phi_j(u_i))_{1 \le i \le n,\ 1 \le j \le p}$ 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 $\min_{1 \le l \le K} \frac{1}{n} X_{B_l}^\top X_{B_l} \succeq h$ such that Condition 1.19 holds with $\tau^2 = h/K = 1/K^2$.

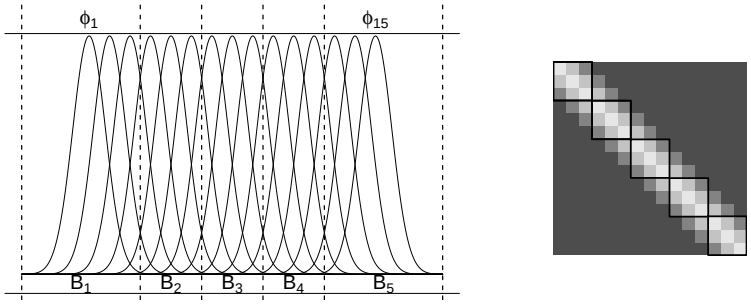


Figure 1.2: Block partitioning of 15 Gaussians into $K = 5$ blocks. The right part shows the corresponding pattern of the Gram matrix.

Applications. As mentioned in the introduction, NNLS has been shown to be remarkably effective in solving deconvolution problems, see §1.5 as well as [96, 98]. 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 above. 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 [5].

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 $n \ge p$, 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

$$\tau_0^2 = \min_{\lambda \in T^{p-1}} \lambda^\top \Sigma \lambda = \rho + \min_{\lambda \in T^{p-1}} (1 - \rho) \|\lambda\|_2^2 = \rho + \frac{1 - \rho}{p}, \tag{1.122}$$

so that the design has the self-regularizing property of Condition 1.19. Let $\emptyset \ne S \subset \{1, \ldots, p\}$ be arbitrary. According to representation (1.82), the corresponding separat-

ing hyperplane constant $\tau^2(S)$ can be evaluated similarly to (1.122). We have

$$\tau^2(S) = \min_{\lambda \in T^{p-s-1}} \lambda^\top \left(\Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}\right)\lambda \tag{1.123}$$

$$= \rho - \rho^2 \mathbf{1}^\top (\Sigma_{SS})^{-1}\mathbf{1} + (1-\rho)\min_{\lambda \in T^{p-s-1}} \|\lambda\|_2^2$$

$$= \rho - \frac{s\rho^2}{1+(s-1)\rho} + \frac{1-\rho}{p-s} = \frac{\rho(1-\rho)}{1+(s-1)\rho} + \frac{1-\rho}{p-s} = \Omega(s^{-1}), \tag{1.124}$$

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. (1.124) 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 (1.124) 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 (1.122) and (1.123), 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 1.37. *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 $z_{(1)} \geq \ldots \geq z_{(p-s)}$ 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:*

$$|F| \overset{\mathcal{D}}{=} s + I\left\{z_{(1)} > 0\right\}\left(1 + \max\left\{1 \leq j \leq p-s-1 : \zeta_j > \theta(s,\rho)\right\}\right), \text{ where}$$
$$\zeta_j = \frac{z_{(j+1)}}{\sum_{k=1}^{j}(z_{(k)} - z_{(j+1)})}, \quad j = 1, \ldots, p-s-1, \quad \text{and } \theta(s,\rho) = \frac{\rho}{1+(s-1)\rho}. \tag{1.125}$$

Proof. We start by noting that Σ is strictly positive definite so that the resulting 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 Theorem 1.28 and 1.29, 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 1.28, we have that

$$\|\widehat{\beta}^{(P1)}\|_1 \leq \frac{2(1+M)\sigma\sqrt{2\log(p)/n}}{\tau^2(S)} \overset{(1.124)}{\leq} \frac{2(1+(s-1)\rho)(1+M)\sigma\sqrt{2\log(p)/n}}{\rho(1-\rho)},$$

with probability at least $1-2p^{-M^2}$, where we have used the closed form expression for $\tau^2(S)$ in (1.124). In order to verify that $\widehat{\beta}_S = \widehat{\beta}^{(P2)} \succ 0$, we follow the back-substitution step (step 2 in the proof of Theorem 1.28) apart from the following modification. In place of (1.97), 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 the proof of Theorem 1.28, 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 1.30 imply that the following block system of inequalities holds.

$$\begin{bmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^cS} & \Sigma_{S^cS^c} \end{bmatrix} \begin{bmatrix} \widehat{\beta}_S \\ \widehat{\beta}_{S^c} \end{bmatrix} \begin{bmatrix} = \\ \preceq \end{bmatrix} \begin{bmatrix} \Sigma_{SS}\beta^*_S + \frac{X_S^\top\varepsilon}{n} \\ \Sigma_{S^cS}\beta^*_S + \frac{X_{S^c}^\top\varepsilon}{n} \end{bmatrix}. \tag{1.126}$$

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.

$$\left(\Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c}\right)\widehat{\beta}_{S^c} \preceq \frac{X_{S^c}^\top(I - X_S(X_S^\top X_S)^{-1}X_S^\top)\varepsilon}{n} = \frac{Z^\top\varepsilon}{n}, \tag{1.127}$$

where $Z = \Pi_S^{\perp} X_{S^c}$ as in (1.80). For equi-correlated design with $\Sigma = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$, we have that

$$\Sigma_{S^cS^c} - \Sigma_{S^cS}(\Sigma_{SS})^{-1}\Sigma_{SS^c} = (1-\rho)I + \underbrace{\frac{\rho(1-\rho)}{1+(s-1)\rho}}_{\gamma(s,\rho)} \mathbf{1}\mathbf{1}^\top = (1-\rho)I + \gamma(s,\rho)\mathbf{1}\mathbf{1}^\top, \quad (1.128)$$

cf. the derivation in (1.124). Denote $\widehat{\alpha} = \widehat{\beta}_{S^c}$, and $G = \{k : \ \widehat{\alpha}_k > 0\}$. Using Lemma 1.30 and (1.128), (1.127) can be written as

$$\begin{aligned} &\frac{Z_k^\top \varepsilon}{n} - (1-\rho)\widehat{\alpha}_k = \gamma(s,\rho)\, \mathbf{1}^\top \widehat{\alpha}, \quad k \in G, \\ &\frac{Z_k^\top \varepsilon}{n} \le \gamma(s,\rho) \mathbf{1}^\top \widehat{\alpha}, \quad k \notin G. \end{aligned} \quad (1.129)$$

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 (1.128), z has the distribution as claimed in Proposition 1.37. From (1.129), we conclude that

$$k \in G \Rightarrow z_k > 0 \ \text{ and } \ z_k \le z_l \text{ for } l \notin G \Rightarrow k \notin G. \quad (1.130)$$

In particular, recalling that $z_{(1)} \ge z_{(2)} \ge \ldots \ge z_{(p-s)}$ denotes the arrangement of the components of z in decreasing order, if $z_{(1)} \le 0$, then (1.129) 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 (1.129) implies that G is non-empty. We may then resolve the first set of equations in (1.129) with respect to $\widehat{\alpha}_G$, which yields

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

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 (1.129), we obtain

$$\begin{aligned} &\frac{\frac{Z_k^\top \varepsilon}{n}}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} - \widehat{\alpha}_k \frac{(1-\rho)+|G|\gamma(s,\rho)}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} = \frac{\gamma(s,\rho)}{1-\rho}, \quad k \in G, \\ &\frac{\frac{Z_k^\top \varepsilon}{n}}{\sum_{\ell \in G}\left(\frac{Z_\ell^\top \varepsilon}{n} - \frac{Z_k^\top \varepsilon}{n}\right)} \le \frac{\gamma(s,\rho)}{1-\rho}, \quad k \notin G. \end{aligned} \quad (1.131)$$

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

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

(1.130) and the inequalities in (1.131), 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 (1.125), 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 $\gamma(0,\rho) = \rho$. □

Proposition 1.37 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 $\frac{1}{2}$. 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 1.3 depicts the $0.01, 0.5, 0.99$-quantiles of the distribution of $|F|$ in virtue of (1.125) 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 1.37 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 [91] and obtain the cardinalities of the active sets. Figure 1.3 shows a strong agreement of the predictions regarding the size of the active set based on the distribution of Proposition 1.37 and the empirical distributions.

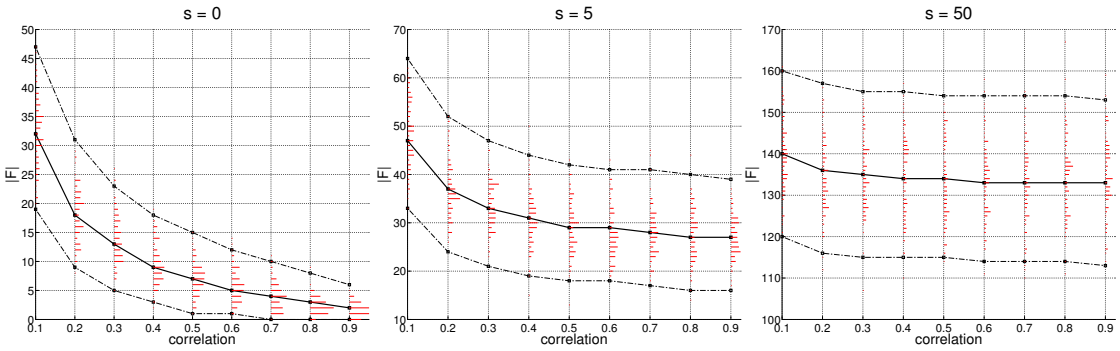


Figure 1.3: Graphical illustration of Proposition 1.37 for $p = 500$. The dotted lines represent the $\{0.01, 0.5, 0.99\}$-quantiles of the distributions obtained from Proposition 1.37 via Monte Carlo simulation. The horizontal bars represent the corresponding relative frequencies based on the solutions of 100 random NNLS problems obtained for each combination of ρ and s.

Random designs. We now reconsider the random matrix ensemble (1.43)

$$\mathrm{Ens}_+(n,p): \; X = (x_{ij})_{\substack{1\le i\le n,\\ 1\le j\le p}}, \; \{x_{ij}\} \text{ i.i.d. from a sub-Gaussian distribution on } \mathbb{R}_+. \tag{1.132}$$

Instances of this class share the property that the population Gram matrix $\Sigma^* = \mathbf{E}[\frac{1}{n}X^\top X]$ equals that of equi-correlated design after rescaling the entries of X by a constant factor. Denoting the mean of the entries and their squares by μ and μ_2, respectively, and setting $\widetilde{X} = X - \mu\mathbb{1}$, where $\mathbb{1}$ denotes a matrix of ones, we have

$$\Sigma^* = \mathbf{E}\left[\frac{1}{n}X^\top X\right] = \mathbf{E}\left[\frac{1}{n}(\widetilde{X}+\mu\mathbb{1})^\top(\widetilde{X}+\mu\mathbb{1})\right] = (\mu_2-\mu^2)I + \mu^2\mathbf{1}\mathbf{1}^\top, \tag{1.133}$$

such that rescaling by $1/\sqrt{\mu_2}$ leads to equi-correlation 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 (1.132) as a basic model for such designs. The study of random matrices is worthwhile particularly because it allows us to address the $n < p$ setting. For this purpose, we will investigate to what extent random matrices from (1.132) inherit properties from the population setting studied in the previous paragraph. As shown in the sequel, the class (1.43) provides instances of designs for which Theorems 1.24 to Theorems 1.29 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 1.24. Recall that Theorem 1.24 requires a combination of the self-regularizing property (Condition 1.19) and the restricted eigenvalue condition (Condition 1.23) 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 [134] on the restricted eigenvalue condition for random matrices with independent sub-Gaussian rows.

Proposition 1.38. *Let X be a random matrix from Ens_+ (1.132) 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 $\delta \in (0,1)$. There exists a constant c_0 depending only on δ and there exist constants $C_1, C_2, c_1, c_2 > 0$*

depending only on δ, ρ and the sub-Gaussian parameter of the entries of X so that if $n \geq C_1 \log p \vee C_2 s \log(c_0\, p/s)$, then, with probability at least $1-\exp(-c_1 n)-2\exp(-c_2 n)$, $\Sigma = X^\top X/n$ has the self-regularizing property with $\tau^2 = \rho/2$ and satisfies condition $\mathsf{RE}(s, 3/\tau^2)$ of Theorem 1.24 with $\phi(s, 3/\tau^2) = (1-\rho)(1-\delta)^2$.

A proof can be found in §1.4.7.

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 (1.124).

Proposition 1.39. *Let X be a random matrix from Ens_+ (1.132) 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 $S \subset \{1,\dots,p\}$, $|S| = s$. Then there exists constants $c, c', C, C' > 0$ depending only on ρ and the sub-Gaussian parameter of the entries of X such that for all $n \geq C s^2 \log(p \vee n)$,*

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

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

A proof can be found in §1.4.7. Proposition 1.39 requires the scaling $n = \Omega(s^2 \log p)$ for $\tau^2(S)$ being positive with high probability for a fixed choice of S, $|S| = s$. This scaling does not match Theorem 1.16, which states that for X as in Proposition 1.39, the associated polyhedral cone $\mathcal{C}_X$ is s-neighbourly, i.e. $\min_{S \in \mathcal{J}(s)} \tau^2(S) > 0$, under the scaling $n = \Omega(s \log(p/s))$. The requirement on the sample size as indicated by Proposition 1.39 turns out to be too strict as confirmed by 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 $S = \{1,\dots,s\}$, 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 [16]. 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: $\{x_{ij}\} \overset{\text{i.i.d.}}{\sim} a\,\mathrm{uniform}([0, \sqrt{3/a}]) + (1-a)\delta_0$, $a \in \{1, \frac{2}{3}, \frac{1}{3}, \frac{2}{15}\}$ $(\rho \in \{\frac{3}{4}, \frac{1}{2}, \frac{1}{3}, \frac{1}{10}\})$

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

We here only display the results for E_1. The results for E_2 to E_4 have been placed into Appendix A; in brief, the results confirm what is shown here. Figure 1.4 displays the 0.05-quantiles of $\tau^2(S)$ over 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 1.39. 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 ($p/n = 10$). This observation is in accordance with Theorem 1.16, which asserts that $\tau^2(S)$ can be positive even with s proportional to n.

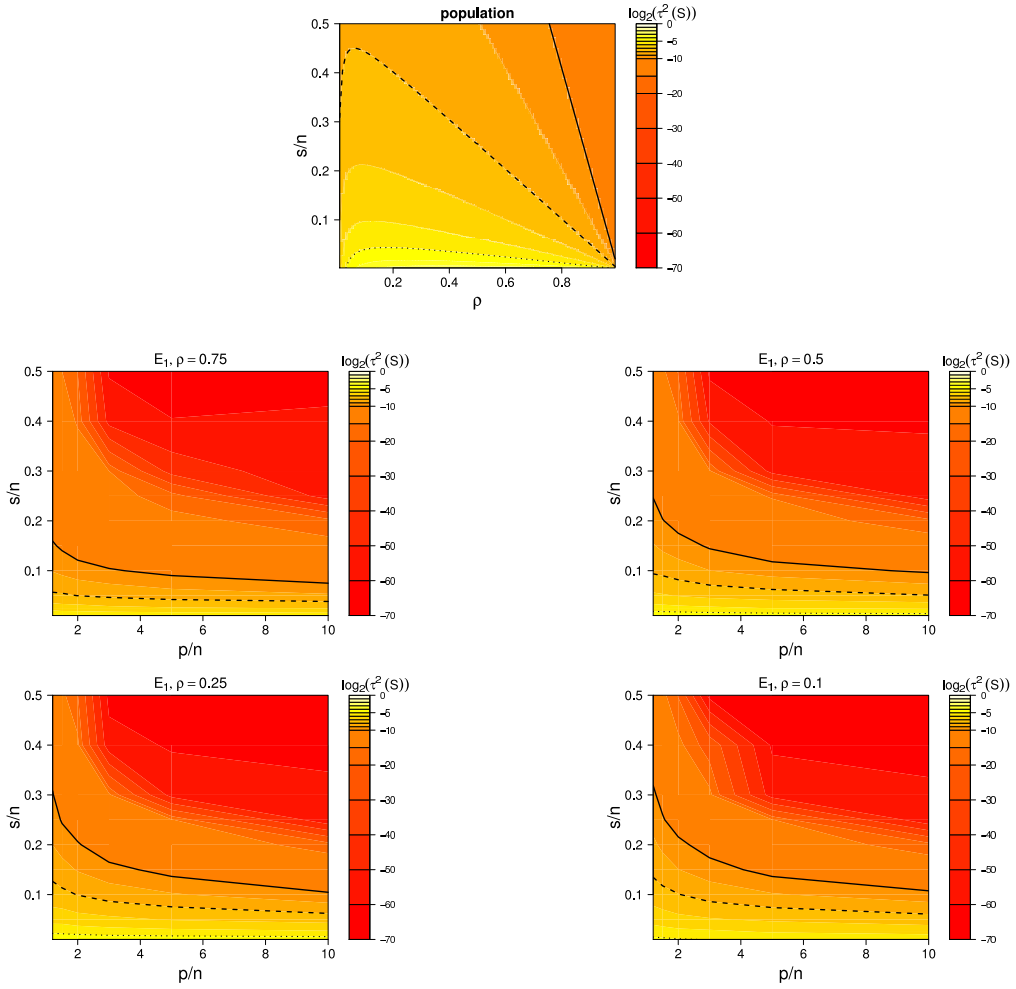


Figure 1.4: Empirical scalings (0.05-quantiles over 100 replications) of the quantity $\log_2(\tau^2(S))$ for random design E_1 from the class Ens_+ in dependency of s/n and p/n, displayed in form of a contour plot. The lines indicate the level set for -10 (solid, $2^{-10} \approx 0.001$), -8 (dashed, $2^{-8} \approx 0.004$) 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 $\rho \in (0, 1)$.

Remark. We point out that Propositions 1.38 and 1.39 can be extended to cover random matrices with i.i.d. rows from a zero-mean Gaussian distribution with equi-correlated components as in Theorem 1.17 as well as to a variant of ensemble $\mathrm{Ens}_{\mathbf{1}}(n, p)$ (1.41) in which the first row is scaled by factor proportional to $\sqrt{n}$. In fact, for X from

$\mathrm{Ens}_{\mathbf{1}}(n,p)$, the population Gram matrix is given by

$$\mathbf{E}\left[\frac{1}{n}X^\top X\right] = \mathbf{E}\left[\frac{1}{n}\widetilde{X}^\top\widetilde{X}\right] + \frac{\mathbf{1}\mathbf{1}^\top}{n}.$$

Since the rows of $\widetilde{X}$ are isotropic, rescaling these rows by $\sqrt{1-\rho}$ and rescaling the row of ones by $\sqrt{\rho n}$ results into equi-correlation.

Sparsity of the solution.
In Proposition 1.37, we have characterized the sparsity in the population setting. It is of interest to investigate this aspect for random design in the $p > n$ setup, particularly in light of Theorem 1.29, 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 1.37. For each combination of $(n, p/n, s/n, \rho)$, 100 replications are considered and the fraction of active variables $|F|/n$ is determined.

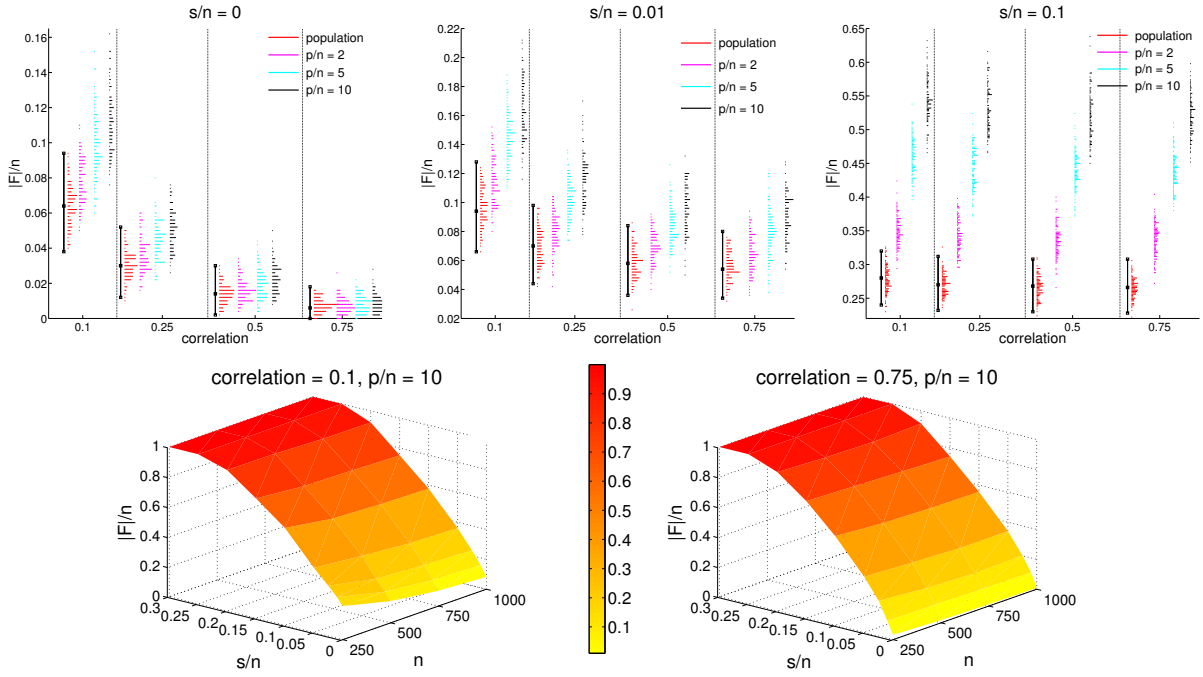


Figure 1.5: Top: sparsity of the NNLS solution for random equi-correlation-like design in the $n < p$ as compared to the population setting. The connected squares represent the 0.01-, 0.5- and 0.99-quantiles of the (conditional) distribution of the fraction of active variables $|F|/n$ in the population setting according to Proposition 1.37. 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.

Figure 1.5 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 1.3). 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 1.33.

1.4.7 Proofs of the results on random matrices

This subsection is devoted to the proofs on several statements on random matrices, in particular Theorems 1.15 to 1.18 and Propositions 1.38 and 1.39.

Restricted eigenvalue properties of random sub-Gaussian matrices with independent rows. Except for the proof of Proposition 1.39, all other proofs rely on Lemma 1.42 below which is a specialized and simplified version of the main result (Theorem 1.6) in [134]. In order to state that result, we need the following preliminaries concerning ψ_2-random variables taken from [34] (see Definition 1.1.1 and Theorem 1.1.5 therein).

Definition 1.40. *A random variable Z is said to be ψ_2 with parameter $\theta > 0$ if*

$$\inf \left\{ a > 0 : \ \mathbf{E}\left[\exp(Z^2/a^2)\right] \leq e \right\} \leq \theta. \tag{1.134}$$

The following lemma establishes a connection between ψ_2 random variables and sub-Gaussian variables.

Lemma 1.41. *[34] If a random variable Z has the property that there exist positive constants C, C' so that $\forall z \geq C'$*

$$\mathbf{P}\left(|Z| \geq z\right) \leq \exp\left(-z^2/C^2\right),$$

then Z is ψ_2 with parameter no more than $2\max(C, C')$.

In view of Lemma 1.41 and the tail bound for zero-mean sub-Gaussian random variables (1.18), it is readily shown that if Z is zero-mean sub-Gaussian with parameter σ, it also holds that Z is ψ_2 with parameter $\theta \leq 4\sigma$.

Lemma 1.42. *[134] Let $\Psi \in \mathbb{R}^{n\times p}$ be a matrix whose rows $\Psi^1, \dots, \Psi^n$, are independent random vectors that are*

(R1) ***isotropic**, i.e. $\mathbf{E}[\langle \Psi^i, u\rangle^2] = 1$ for every unit vector $u \in \mathbb{R}^p$, $i = 1, \dots, n$,*

(R2) ***ψ_2**, i.e. there exists $\theta > 0$ such that for every unit vector $u \in \mathbb{R}^p$*

$$\inf\left\{a > 0 : \mathbf{E}\left[\exp(\langle \Psi^i, u\rangle^2 / a^2)\right] \le e\right\} \le \theta, \quad i = 1, \dots, n. \tag{1.135}$$

Let further $R \in \mathbb{R}^{p\times p}$ be a positive definite matrix with minimum eigenvalue $\vartheta > 0$ and set $\Gamma = \frac{1}{n} R^\top \Psi^\top \Psi R$. Then, for any $\delta \in (0,1)$ and any $\alpha \in [1, \infty)$, there exist positive constants C_θ, $c > 0$ (the first depending on the ψ_2 parameter θ) so that if

$$n \ge \frac{C_\theta}{\delta^2} s \left(1 + \frac{16(3\alpha^2)(3\alpha+1)}{\vartheta^2 \delta^2}\right) \log\left(c \frac{p}{s\delta}\right),$$

with probability at least $1 - 2\exp(-\delta^2 n / C_\theta)$, Γ satisfies condition $\mathsf{RE}(s, \alpha)$ with $\phi(s, \alpha) = \vartheta^2(1-\delta)^2$.

Proofs of Theorem 1.15 to 1.18 and Proposition 1.38. To prove Theorems 1.15 to 1.18, we proceed according to the scheme following Theorem 1.13. We restate that scheme for convenience.

Verify.$(\mathcal{H})$:	Verify whether X satisfies condition $(\mathcal{H})$.
Bound.η_h:	If this is the case, upper bound $\eta_h = \max_{1\le j\le p} h_j \big/ \min_{1\le j\le p} h_j$, where $h = X^\top w \succ 0$. The smaller η_h, the easier the next step.
Verify.$\mathsf{RN}(s, \eta_h)$:	Verify whether X satisfies condition $\mathsf{RN}(s, \eta_h)$.

Recall that if one succeeds in the last step, $\mathcal{C}_X$ is s-neighbourly.
Let us now proceed according to the above steps for ensemble $\mathrm{Ens}_{\mathbf{1}}(n, p)$.

Proof. (Theorem 1.15) Step *Bound.*η_h can be performed easily as $X^\top e_1 = \mathbf{1}$ and thus $\eta_h = 1$, where e_1 denotes the first canonical unit vector. Noting that $\mathcal{N}(X) \subseteq \mathcal{N}(\widetilde{X})$, it remains to check whether $\widetilde{X}$ satisfies condition $\mathsf{RN}(s, 1)$ with s of the order (1.42). For this purpose, we want to apply Lemma 1.42 with $\Psi = \widetilde{X}$, $R = I_p$ and $\alpha = 1$, which would yield that condition $\mathsf{RE}(s, 1)$ (and thus in particular condition $\mathsf{RN}(s, 1)$) are satisfied once $n - 1 > C_1 s \log(C_2 p/s)$ as asserted in the theorem. It remains to verify whether $\widetilde{X}$ has the properties (R1) and (R2). The former holds by construction, while the latter follows from the fact that for any unit vector u, the random variables $\langle \widetilde{X}^i, u\rangle$, $i = 1, \dots, n$, are zero-mean sub-Gaussian and hence also ψ_2 according to the remark following Lemma 1.41. □

Proof. (Theorem 1.18) By rotational invariance of the Gaussian distribution, we may fix $w = e_1$ in Construction $\mathtt{HALFSPACE}(n, p, t)$ so that $X = [h^\top;\ \widetilde{X}]$, where the entries of h are i.i.d. from a t-truncated Gaussian distribution and the entries of $\widetilde{X}$ are

i.i.d. standard Gaussian. Concerning step *Bound.*η_h, we have by construction that $\eta_h \le t^{-1}\max_{1\le j\le p} h_j$. For $z \ge 0$[9], we have the tail bound

$$\begin{aligned}\mathbf{P}(h_1 > z) &\le C_t \int_z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du, \qquad C_t := \frac{1}{1 - \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du} \\ &\le C_t \exp\left(-\frac{z^2}{2}\right)\end{aligned}$$

By the union bound, we obtain that

$$\mathbf{P}\left(\eta_h > \sqrt{\frac{4\log p}{t^2}}\right) \le C_t/p.$$

We work on conditional on the event $\{\eta_h \le \sqrt{4\log(p)/t^2}\}$. As in the previous proof, we use that $\mathcal{N}(X) \subseteq \mathcal{N}(\widetilde{X})$ and invoke Lemma 1.42 with $\Psi = \widetilde{X}$ and $R = I_p$, but this time with $\alpha = \sqrt{4\log(p)/t^2}$ so that we deduce (1.45) with the probability as stated. □

Remark. In Example 2 at the beginning of §1.4.6, it is claimed that for X generated according $\texttt{HALFSPACE}(n, p, t)$, we have $\tau_0^2 = O_{\mathbf{P}}(1/n)$ once $p > 2n$, i.e. X does not have the self-regularizing property. An upper bound on τ_0^2 can be established by noting that $\widetilde{X}$ in the above proof belongs to $\mathrm{Ens}_0(n-1, p)$ (1.38). In light of Wendel's Theorem and (1.40), with high probability, there exists $\lambda_0 \in T^{p-1}$ such that $\|\widetilde{X}\lambda_0\|_2^2 = 0$. Accordingly, we have

$$\tau_0^2 = \min_{\lambda\in T^{p-1}} \left\{ \frac{\langle h, \lambda\rangle^2}{n} + \frac{1}{n}\|\widetilde{X}\lambda\|_2^2 \right\} \le \frac{\langle h, \lambda_0\rangle^2}{n} + \frac{1}{n}\|\widetilde{X}\lambda_0\|_2^2 = O_{\mathbf{P}}(1/n),$$

where we have used that $\langle h, \lambda_0\rangle = O_{\mathbf{P}}(\|\lambda_0\|_2) = O_{\mathbf{P}}(\|\lambda_0\|_1) = O_{\mathbf{P}}(1)$, which follows from the fact that λ_0 is independent of h and the fact that the entries of h are i.i.d. sub-Gaussian.

Proof. (Theorem 1.17) We first bound η_h. Let us write $X = \Psi(\Sigma^*)^{1/2}$, with $(\Sigma^*)^{1/2}$ denoting the symmetric root of $\Sigma^* = (1-\rho)I_p + \rho\mathbf{1}\mathbf{1}^\top$. The largest eigenvalue of Σ^* is given by $\phi_1 = 1 + (p-1)\rho$ with corresponding eigenvector $u_1 = \mathbf{1}/\sqrt{p}$. The remaining eigenvalues $\phi_2, \ldots, \phi_p$ are all equal to $1-\rho$, and we denote the corresponding eigenvectors by $u_2, \ldots, u_p$. Take w as the vectors of signs of $\Psi\mathbf{1}$ and set $h = X^\top w$. We then have

$$\begin{aligned} h &= (\Sigma^*)^{1/2}\Psi^\top w \\ &= \sqrt{\phi_1}\frac{\mathbf{1}\mathbf{1}^\top}{p}\Psi^\top w + \sum_{j=2}^p \sqrt{\phi_j} u_j u_j^\top \Psi^\top w \\ &= c_\rho \mathbf{1}\frac{(\Psi\mathbf{1})^\top w}{\sqrt{p}} + (\Sigma^*)^\perp \Psi^\top w, \end{aligned} \tag{1.136}$$

[9]Note that for $z \in [0, t]$, the given bound exceeds one and hence holds trivially

where

$$c_\rho = \sqrt{\frac{\phi_1}{p}} = \sqrt{\frac{1+(p-1)\rho}{p}}, \quad \text{and} \quad (\Sigma^*)^\perp = \sum_{j=2}^{p} \sqrt{\phi_j} u_j u_j^\top.$$

Now note that by construction of w

$$Z = (\Psi \mathbf{1}/\sqrt{p})^\top w = \sum_{i=1}^{n} \left| \sum_{j=1}^{p} \Psi_{ij}/\sqrt{p} \right| = \sum_{i=1}^{n} \zeta_i, \tag{1.137}$$

where the $\{\zeta_i\}_{i=1}^n$ are i.i.d. half-Gaussian random variables. Regarding the second term in (1.136), let

$$\xi_j = e_j^\top (\Sigma^*)^\perp \Psi^\top w, \;\; j = 1, \ldots, p.$$

For $j = 1, \ldots, p$ and $t \geq 0$, we have

$$\mathbf{P}\left(\xi_j > t\right) = \mathbf{E}_w \left[\mathbf{P}\left(\xi_j > t | w = \sigma\right) \right],$$

where $\mathbf{E}_w$ denotes expectation w.r.t. w, whose distribution is uniform over $\{-1, 1\}^n$. Let g be a p-dimensional standard Gaussian random vector. Working conditional on σ, we have for $j = 1, \ldots, p$

$$\begin{aligned}
\mathbf{P}\left(\xi_j > t | w = \sigma\right) &= \mathbf{P}\left(e_j^\top (\Sigma^*)^\perp \Psi^\top w > t | w = \sigma\right) \\
&= \mathbf{P}\left(e_j^\top (\Sigma^*)^\perp \left\|w\right\|_2 g > t\right) \\
&= \mathbf{P}\left(e_j^\top (\Sigma^*)^\perp \sqrt{n} g > t\right), \\
&\leq \mathbf{P}(\sqrt{1-\rho}\sqrt{n} g_1 > t)
\end{aligned}$$

Using (1.19) with $\mathbf{Z} = \xi$, $v_j = e_j$, $j = 1, \ldots, p$, and $z = \sqrt{2 \log p}$, we obtain

$$\mathbf{P}\left(\max_{1 \leq j \leq p} |\xi_j| > 2\sqrt{2 \log p}\sqrt{1-\rho}\sqrt{n} \right) \leq 2/p. \tag{1.138}$$

We now derive a lower bound on Z given in (1.137). Denote by $\mu = (2/\pi)^{1/2}$ the mean of the $\{\zeta_i\}_{i=1}^n$ and let $\widetilde{\zeta}_i = \min\{\zeta_i, 2\mu\}$ so that $\mathbf{E}[\widetilde{\zeta}_i] \geq 3/4\mu$, $i = 1, \ldots, n$. We hence have

$$\mathbf{P}\left(Z < \frac{1}{2} n\mu\right) \leq \mathbf{P}\left(\sum_{i=1}^{n} \widetilde{\zeta}_i < \frac{1}{2} n\mu\right) \leq \mathbf{P}\left(\sum_{i=1}^{n} \left(\widetilde{\zeta}_i - \mathbf{E}[\widetilde{\zeta}_i]\right) < -\frac{1}{4} n\mu\right) \leq \exp\left(-\frac{n}{32}\right) \tag{1.139}$$

where the last inequality is an application of Hoeffding's inequality. Combining (1.136) to (1.139), we have with probability at least $1 - \exp(-n/32) - 2/p$

$$\eta_h = \frac{\max_{1 \leq j \leq p} h_j}{\min_{1 \leq j \leq p} h_j} \leq \frac{\frac{1}{2} c_\rho n\mu + 2\sqrt{2 \log p}\sqrt{n}}{\frac{1}{2} c_\rho n\mu - 2\sqrt{2 \log p}\sqrt{n}}.$$

It follows that $\eta_h \leq 3$ (the constant 3 is chosen for convenience here) if

$$2\sqrt{2 \log p}\sqrt{n} \leq \frac{1}{4} c_\rho n\mu \iff n \geq \frac{128}{\mu^2 c_\rho^2} \log p =: C_0 \log p. \tag{1.140}$$

To finish the proof, we apply Lemma 1.42 conditional on the event $\{\eta_h \leq 3\}$ with Ψ as a $n \times p$ random standard Gaussian matrix, $R = (\Sigma^*)^{1/2}$ and $\alpha = 3$. We conclude that if n satisfies (1.140), $\mathcal{C}_X$ is s-neighbourly as long as $s \leq n/(C_1 \log(C_2 \frac{p}{s}))$, with probability at least $1 - 3\exp(-cn) - 2/p$. □

Proof. (Theorem 1.16) Let us recall the decomposition of X into a centered matrix $\widetilde{X}$ and a matrix with constant entries μ equal to the mean of the entries of X

$$X = \widetilde{X} + \mu \mathbb{1}, \tag{1.141}$$

as already used for (1.133). Let us assume without loss of generality that the entries of X are scaled such that $\mu_2 := \mathbf{E}[x_{11}^2] = 1$. As explained below (1.133), $\Sigma^* := \mathbf{E}[\frac{1}{n} X^\top X]$ then has equi-correlation structure with $\rho = \mu^2 \in (0, 1)$. Accordingly, decomposition (1.141) becomes

$$X = \widetilde{X} + \sqrt{\rho}\mathbb{1}, \tag{1.142}$$

Let further denote σ the sub-Gaussian parameter of the entries of X. For step *Bound.*η_h, we choose $w = \mathbf{1}$ so that $h = X^\top \mathbf{1} = \widetilde{X}^\top \mathbf{1} + n\sqrt{\rho}\mathbf{1}$. The entries of $\widetilde{X}^\top \mathbf{1}$ are i.i.d. zero-mean sub-Gaussian with parameter $\sigma n^{1/2}$. The sub-Gaussian tail bound (1.18) implies that for any $\varepsilon \in (0, 1)$

$$\mathbf{P}(h_1 \notin [(1-\varepsilon)n\sqrt{\rho}, (1+\varepsilon)n\sqrt{\rho}]) \leq 2\exp\left(-\frac{n\varepsilon^2\rho}{2\sigma^2}\right).$$

Consequently, if $n \geq 8\sigma^2(\varepsilon\sqrt{\rho})^{-2}\log p =: C_0 \log p$, the union bound yields that

$$\mathbf{P}(\exists j \in \{1, \ldots, p\} : h_j \notin [(1-\varepsilon)n\sqrt{\rho}, (1+\varepsilon)n\sqrt{\rho}]) \leq \exp\left(-\frac{n\varepsilon^2\rho}{4\sigma^2}\right). \tag{1.143}$$

In the sequel, it will be shown that Lemma 1.42 can be applied with $\Psi = X(\Sigma^*)^{-1/2}$ and $R = (\Sigma^*)^{1/2}$ and $\alpha = (1+\varepsilon)/(1-\varepsilon)$ for an arbitrary choice of ε between 0 and 1. The rows of Ψ are isotropic by construction, so that requirement (R1) in Lemma 1.42 is fulfilled. In the remainder, we verify that the rows of Ψ fulfill requirement (R2) in Lemma 1.42. Since the rows of Ψ are i.i.d., it suffices to consider a single row. Write X^1 for the transpose of the first row of X and accordingly $\widetilde{X}^1$ for that of $\widetilde{X}$. We have for any unit vector $u \in \mathbb{R}^p$

$$\begin{aligned}
\left\langle \Psi^1, u \right\rangle = \left\langle (\Sigma^*)^{-1/2} X^1, u \right\rangle &= \left\langle (\Sigma^*)^{-1/2} (\widetilde{X}^1 + \sqrt{\rho}\mathbf{1}), u \right\rangle \\
&= \left\langle \widetilde{X}^1, (\Sigma^*)^{-1/2} u \right\rangle + \sqrt{\frac{\rho}{(1-\rho) + p\rho}} \langle \mathbf{1}, u \rangle \\
&\leq \left\langle \widetilde{X}^1, (\Sigma^*)^{-1/2} u \right\rangle + 1.
\end{aligned}$$

For the second equality, we have used that $\mathbf{1}$ is an eigenvector of Σ^* with eigenvalue $1 + (p-1)\rho$, while the inequality results from Cauchy-Schwarz. We now estimate the moment-generating function of the random variable $\langle \Psi^1, u \rangle$ as follows. For any $t \geq 0$,

we have

$$\begin{aligned}\mathbf{E}[\exp(t\left\langle \Psi^1, u\right\rangle)] &\leq \exp(t)\,\mathbf{E}\left[\exp\left(t\left\langle \widetilde{X}^1, (\Sigma^*)^{-1/2}\,u\right\rangle\right)\right] \\ &\leq \exp(t)\,\mathbf{E}\left[\exp\left(\frac{\sigma^2 t^2}{2}\|(\Sigma^*)^{-1/2}\,u\|_2^2\right)\right] \\ &\leq \exp(t)\exp\left(\frac{\sigma^2 t^2}{2(1-\rho)}\right) \\ &\leq e\exp\left(\frac{(\sigma^2+2)t^2}{2(1-\rho)}\right) = e\exp\left(\frac{\widetilde{\sigma}^2 t^2}{2}\right),\end{aligned}$$

where $\widetilde{\sigma} = \sqrt{(\sigma^2+2)/(1-\rho)}$. For the third equality, we have used that the maximum eigenvalue of $(\Sigma^*)^{-1}$ equals $(1-\rho)^{-1}$. Analogously, we obtain that

$$-\left\langle \Psi^1, u\right\rangle \leq \left\langle -\widetilde{X}^1, (\Sigma^*)^{-1/2}\,u\right\rangle + 1, \quad \text{and} \quad \mathbf{E}[\exp(t\left\langle -\Psi^1, u\right\rangle)] \leq e\exp\left(\frac{\widetilde{\sigma}^2 t^2}{2}\right)\ \forall t \geq 0.$$

From the Chernov method, we hence obtain that for any $z \geq 0$

$$\mathbf{P}(|\left\langle \Psi^1, u\right\rangle| > z) \leq 2e\exp\left(-\frac{z^2}{2\widetilde{\sigma}^2}\right).$$

Invoking Lemma 1.41 with $C' = \widetilde{\sigma}\sqrt{3\log(2e)}$ and $C = \sqrt{6}\widetilde{\sigma}$, it follows that the random variable $\langle \Psi^1, u\rangle$ is ψ_2 with parameter $2\sqrt{6}\widetilde{\sigma}$, and we conclude that the rows of the matrix Ψ indeed satisfy condition (1.135) with θ equal to that value of the parameter. □

The proof of Proposition 1.38 is along the lines of the previous proof.

Proof. (Proposition 1.38) We first show that Σ satisfies the self-regularizing property with $\tau^2 \geq \rho/2$ with probability at least $1-\exp(-c_{\sigma,\rho}n)$ as long as $n > C_{\sigma,\rho}\log p$, where $C_{\sigma,\rho}$ and $c_{\sigma,\rho}$ are positive constants only depending on the sub-Gaussian parameter σ and ρ. By definition of τ_0 (1.46), we have

$$\tau_0 = \max_{\tau, w} \tau \quad \text{subject to } n^{-1/2}X^\top w \succeq \tau\mathbf{1}, \quad \|w\|_2 \leq 1. \tag{1.144}$$

We now choose $w = \mathbf{1}/\sqrt{n}$ and $\tau = \sqrt{\rho}(1-\varepsilon)$, where $\varepsilon = 1 - 1/\sqrt{2}$. Arguing as for (1.143), we have that these choices are feasible for (1.144) with probability at least $1-\exp(-n\rho(\sqrt{2}-1)^2/(8\sigma^2))$ as long as $n > 16(\sqrt{2}-1)^{-2}\rho^{-1}\sigma^2\log p$. Conditional on the event $\{\tau^2 \geq \rho/2\}$, we invoke Lemma 1.42 with $\Psi = X(\Sigma^*)^{-1/2}$, $R = (\Sigma^*)^{1/2}$, $\vartheta^2 = 1-\rho$ and $\alpha = 3/\tau^2 \leq 6/\rho$ as is justified given the previous proof. □

Proof of Proposition 1.39. The proof of Proposition 1.39 relies on several lemmas which are stated below.

Bernstein-type inequality for squared sub-Gaussian random variables. The following tail bound results from Lemma 14, Proposition 16 and Remark 18 in [161].

Lemma 1.43. *Let $Z_1, \dots, Z_m$ be i.i.d. zero-mean sub-Gaussian random variables with parameter σ and the property that $\mathbf{E}[Z_1^2] \le 1$. Then for any $z \ge 0$, one has*

$$\mathbf{P}\left(\sum_{i=1}^m Z_i^2 > m + zm\right) \le \exp\left(-c \min\left\{\frac{z^2}{\sigma^4}, \frac{z}{\sigma^2}\right\} m\right), \tag{1.145}$$

where $c > 0$ is an absolute constant.

Concentration of extreme singular values of sub-Gaussian random matrices. Let $s_{\min}(A)$ and $s_{\max}(A)$ denote the minimum and maximum singular value of a matrix A. The following lemma is a special case of Theorem 39 in [161].

Lemma 1.44. *Let A be an $n \times s$ matrix with i.i.d. zero-mean sub-Gaussian entries with sub-Gaussian parameter σ and unit variance. Then for every $z \ge 0$, with probability at least $1 - 2\exp(-cz^2)$, one has*

$$s_{\max}\left(\frac{1}{n}A^\top A - I\right) \le \max(\delta, \delta^2), \quad \text{where } \delta = C\sqrt{\frac{s}{n}} + \frac{z}{\sqrt{n}}, \tag{1.146}$$

with C, c depending only on σ.

Entry-wise concentration of the Gram matrix associated with a sub-Gaussian random matrix. The next lemma results from Lemma 1 in [125] and the union bound.

Lemma 1.45. *Let X be an $n \times p$ random matrix of i.i.d. zero-mean, unit variance sub-Gaussian entries with parameter σ. Then*

$$\mathbf{P}\left(\max_{1 \le j,k \le p} \left|\left(\frac{1}{n}X^\top X - I\right)_{jk}\right| > z\right) \le 4p^2 \exp\left(-\frac{nz^2}{128(1+4\sigma^2)^2}\right) \tag{1.147}$$

for all $z \in (0, 8(1+4\sigma^2))$.

Equipped with these results, we are in position to derive several intermediate results to be used in the proof of Proposition 1.39. We first prove an upper bound on the entry-wise difference between Σ and its population counterpart Σ^* for X from Ens_+. Recalling decomposition (1.142), we have the following expansion for Σ.

$$\Sigma = \frac{1}{n}\widetilde{X}^\top \widetilde{X} + \sqrt{\rho}\left(\frac{1}{n}\widetilde{X}^\top \mathbb{1} + \frac{1}{n}\mathbb{1}^\top \widetilde{X}\right) + \rho \mathbf{1}\mathbf{1}^\top, \quad \text{where } \mathbf{E}\left[\frac{1}{n}\widetilde{X}^\top \widetilde{X}\right] = (1-\rho)I. \tag{1.148}$$

Observe that

$$n^{-1}\widetilde{X}^\top \mathbb{1} = D\mathbf{1}\mathbf{1}^\top, \quad \text{and } n^{-1}\mathbb{1}^\top \widetilde{X} = \mathbf{1}\mathbf{1}^\top D, \tag{1.149}$$

where $D \in \mathbb{R}^{p \times p}$ is a diagonal matrix with diagonal entries $d_{jj} = n^{-1}\sum_{i=1}^n \widetilde{x}_{ij}$, $j = 1, \dots, p$. It hence follows from (1.19) that

$$\mathbf{P}\left(\max_{j,k}\left|n^{-1}\widetilde{X}^\top \mathbb{1}\right|_{jk} > 2\sigma\sqrt{\frac{2\log(p \vee n)}{n}}\right) \le \frac{2}{p \vee n}, \tag{1.150}$$

Combining (1.148), (1.150) and Lemma 1.45, it follows that there exists a constant $C > 0$ depending only on σ such that

$$\mathbf{P}\left(\max_{j,k}\left|(\Sigma - \Sigma^*)_{jk}\right| > C\sqrt{\frac{\log(p \vee n)}{n}}\right) \leq \frac{6}{p \vee n}. \tag{1.151}$$

Let now $S \subset \{1, \ldots, p\}$, $|S| = s < n$ be given. Without loss of generality, let us assume that $S = \{1, \ldots, s\}$. In the sequel, we control $s_{\max}(\Sigma^*_{SS} - \Sigma_{SS})$. From decomposition (1.148), we obtain that

$$s_{\max}(\Sigma^*_{SS} - \Sigma_{SS}) \leq (1-\rho)s_{\max}\left(\frac{1}{1-\rho}\frac{\widetilde{X}_S^\top \widetilde{X}_S}{n} - I\right) + 2\sqrt{\rho}s_{\max}\left(\frac{\widetilde{X}_S^\top \mathbb{1}_S}{n}\right) \tag{1.152}$$

Introduce $w = (\sum_{i=1}^n \widetilde{x}_{i1}/n, \ldots, \sum_{i=1}^n \widetilde{x}_{is}/n)^\top$ as the vector of column means of $\widetilde{X}_S$. We have that

$$s_{\max}\left(\frac{\widetilde{X}_S^\top \mathbb{1}_S}{n}\right) = \sup_{\|u\|_2=1}\sup_{\|v\|_2=1} u^\top \frac{\widetilde{X}_S^\top \mathbb{1}_S}{n} v = \sup_{\|u\|_2=1}\sup_{\|v\|_2=1} u^\top w \mathbf{1}^\top v = \sqrt{s}\|w\|_2. \tag{1.153}$$

Moreover,

$$\|w\|_2^2 = \sum_{j=1}^s \left(\frac{\sum_{i=1}^n \widetilde{x}_{ij}}{n}\right)^2 = \frac{1}{n}\sum_{j=1}^s z_j^2, \quad \text{where } z_j = n^{-1/2}\sum_{i=1}^n \widetilde{x}_{ij}. \tag{1.154}$$

Noting that the $\{z_j\}_{j=1}^s$ are i.i.d. zero-mean sub-Gaussian random variables with parameter σ and variance no larger than one, we are in position to apply Lemma 1.43, which yields that for any $t \geq 0$

$$\mathbf{P}\left(\|w\|_2^2 > \frac{s}{n}(1+t)\right) \leq \exp\left(-c\min\left(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right)s\right). \tag{1.155}$$

Combining (1.152), (1.153) and (1.155) and using Lemma 1.44 to control the term $s_{\max}\left(\frac{1}{1-\rho}\frac{\widetilde{X}_S^\top \widetilde{X}_S}{n} - I\right)$, we obtain that for any $t \geq 0$ and any $z \geq 0$

$$\begin{aligned}&\mathbf{P}\left(s_{\max}(\Sigma^*_{SS} - \Sigma_{SS}) > \max\left\{C\sqrt{\frac{s}{n}} + \frac{z}{\sqrt{n}}, \left(C\sqrt{\frac{s}{n}} + \frac{z}{\sqrt{n}}\right)^2\right\} + 2\sqrt{\frac{s^2(1+t)}{n}}\right) \\ &\leq \exp(-c_1\min\{t, t^2\}s) - 2\exp(-c_2 z^2),\end{aligned} \tag{1.156}$$

where $C, c_1, c_2 > 0$ only depend on the sub-Gaussian parameter σ.

Proof. (Proposition 1.39) The scaling of $\tau^2(S)$ is analyzed based on representation (1.81)

$$\tau^2(S) = \min_{\theta \in \mathbb{R}^s,\, \lambda \in T^{p-s-1}} \frac{1}{n}\left\|X_S\theta - X_{S^c}\lambda\right\|_2^2. \tag{1.157}$$

In the following, denote by $\mathbb{S}^{s-1} = \{u \in \mathbb{R}^s : \|u\|_2 = 1\}$ the unit sphere in $\mathbb{R}^s$. Expanding the square in (1.157), we have

$$\begin{aligned}
\tau^2(S) &= \min_{\theta\in\mathbb{R}^s,\, \lambda\in T^{p-s-1}} \theta^\top \Sigma_{SS}\theta - 2\theta^\top\Sigma_{SS^c}\lambda + \lambda^\top\Sigma_{S^cS^c}\lambda \\
&\geq \min_{r\geq 0,\, u\in\mathbb{S}^{s-1},\, \lambda\in T^{p-s-1}} r^2 u^\top\Sigma^*_{SS}u - r^2 s_{\max}\left(\Sigma_{SS} - \Sigma^*_{SS}\right) - \\
&\quad - 2ru^\top\Sigma_{SS^c}\lambda + \lambda^\top\Sigma_{S^cS^c}\lambda \\
&\geq \min_{r\geq 0,\, u\in\mathbb{S}^{s-1},\, \lambda\in T^{p-s-1}} r^2 u^\top\Sigma^*_{SS}u - r^2 s_{\max}\left(\Sigma_{SS} - \Sigma^*_{SS}\right) \\
&\quad - 2\rho ru^\top\mathbf{1} - 2ru^\top(\Sigma_{SS^c} - \Sigma^*_{SS^c})\lambda + \rho + \frac{1-\rho}{p-s} - \\
&\quad - \max_{\lambda\in T^{p-s-1}} \left|\lambda^\top(\Sigma_{S^cS^c} - \Sigma^*_{S^cS^c})\lambda\right|.
\end{aligned} \tag{1.158}$$

For the last inequality, we have used that $\min_{\lambda\in T^{p-s-1}} \lambda^\top\Sigma^*_{S^cS^c}\lambda = \rho + \frac{1-\rho}{p-s}$. We further set

$$\Delta = s_{\max}\left(\Sigma_{SS} - \Sigma^*_{SS}\right), \tag{1.159}$$

$$\delta = \max_{u\in\mathbb{S}^{s-1},\lambda\in T^{p-s-1}} \left|u^\top\left(\Sigma_{S^cS^c} - \Sigma^*_{S^cS^c}\right)\lambda\right|. \tag{1.160}$$

The random terms Δ and δ will be controlled uniformly over $u \in \mathbb{S}^{s-1}$ and $\lambda \in T^{p-s-1}$ below by invoking (1.151) and (1.156). For the moment, we treat these two terms as constants. We now minimize the lower bound in (1.158) w.r.t. u and r separately from λ. This minimization problem involving u and r only reads

$$\min_{r\geq 0,\, u\in\mathbb{S}^{s-1}} r^2 u^\top\Sigma^*_{SS}u - 2\rho ru^\top\mathbf{1} - r^2\Delta - 2r\delta. \tag{1.161}$$

We first derive an expression for

$$\phi(r) = \min_{u\in\mathbb{S}^{s-1}} r^2 u^\top\Sigma^*_{SS}u - 2\rho ru^\top\mathbf{1}. \tag{1.162}$$

We decompose $u = u^{\parallel} + u^{\perp}$, where $u^{\parallel} = \left\langle \frac{\mathbf{1}}{\sqrt{s}}, u\right\rangle \frac{\mathbf{1}}{\sqrt{s}}$ is the projection of u on the unit vector $\mathbf{1}/\sqrt{s}$, which is an eigenvector of Σ^*_{SS} associated with its largest eigenvalue $1 + \rho(s-1)$. By Parseval's identity, we have $\|u^{\parallel}\|_2^2 = \gamma$, $\|u^{\perp}\|_2^2 = (1-\gamma)$ for some $\gamma \in [0,1]$. Inserting this decomposition into (1.162) and noting that the remaining eigenvalues of Σ^*_{SS} are all equal to $(1-\rho)$, we obtain that

$$\begin{aligned}
&\phi(r) = \min_{\gamma\in[0,1]} \Phi(\gamma, r), \\
&\text{with } \Phi(\gamma, r) = r^2\gamma \underbrace{(1 + (s-1)\rho)}_{s_{\max}(\Sigma^*_{SS})} + r^2(1-\gamma)\underbrace{(1-\rho)}_{s_{\min}(\Sigma^*_{SS})} - 2\rho r\sqrt{\gamma}\sqrt{s},
\end{aligned} \tag{1.163}$$

where we have used that $\langle u^{\perp}, \mathbf{1}\rangle = 0$. Let us put aside the constraint $\gamma \in [0,1]$ for a moment. The function Φ in (1.163) is a convex function of γ, hence we may find an (unconstrained) minimizer $\widetilde{\gamma}$ by differentiating and setting the derivative equal to

zero. This yields $\widetilde{\gamma} = \frac{1}{r^2 s}$, which coincides with the constrained minimizer if and only if $r \geq \frac{1}{\sqrt{s}}$. Otherwise, $\widetilde{\gamma} \in \{0, 1\}$. We can rule out the case $\widetilde{\gamma} = 0$, since for all $r < 1/\sqrt{s}$

$$\Phi(0, r) = r^2(1-\rho) > r^2(1 + (s-1)\rho) - 2\rho r\sqrt{s} = \Phi(1, r).$$

We have $\Phi(\frac{1}{r^2 s}, r) = r^2(1-\rho) - \rho$ and $\Phi(\frac{1}{r^2 s}, \frac{1}{\sqrt{s}}) = \Phi(1, \frac{1}{\sqrt{s}})$. Hence, the function $\phi(r)$ in (1.162) is given by

$$\phi(r) = \begin{cases} r^2 s_{\max}(\Sigma^*_{SS}) - 2\rho r\sqrt{s} & r \leq 1/\sqrt{s}, \\ r^2(1-\rho) - \rho & \text{otherwise.} \end{cases} \tag{1.164}$$

The minimization problem (1.161) to be considered eventually reads

$$\min_{r \geq 0} \psi(r), \quad \text{where } \psi(r) = \phi(r) - r^2\Delta - 2r\delta. \tag{1.165}$$

We argue that it suffices to consider the case $r \leq 1/\sqrt{s}$ in (1.164) provided

$$((1-\rho) - \Delta) > \delta\sqrt{s}, \tag{1.166}$$

a condition we will comment on below. If this condition is met, differentiating shows that ψ is increasing on $(\frac{1}{\sqrt{s}}, \infty)$. In fact, for all r in that interval,

$$\frac{d}{dr}\psi(r) = 2r(1-\rho) - 2r\Delta - 2\delta, \text{ and thus}$$

$$\frac{d}{dr}\psi(r) > 0 \text{ for all } r \in \left(\frac{1}{\sqrt{s}}, \infty\right) \Leftrightarrow \frac{1}{\sqrt{s}}((1-\rho) - \Delta) > \delta.$$

Considering the case $r \leq 1/\sqrt{s}$, we observe that $\psi(r)$ is convex provided

$$s_{\max}(\Sigma^*_{SS}) > \Delta, \tag{1.167}$$

a condition we shall comment on below as well. Provided (1.166) and (1.167) hold true, differentiating (1.165) and setting the result equal to zero, we obtain that the minimizer $\widehat{r}$ of (1.165) is given by $(\rho\sqrt{s} + \delta)/(s_{\max}(\Sigma^*_{SS}) - \Delta)$. Substituting this result back into (1.165) and in turn into the lower bound (1.158), one obtains after collecting terms

$$\begin{aligned} \tau^2(S) \geq & \rho\frac{(1-\rho) - \Delta}{(1-\rho) + s\rho - \Delta} - \frac{2\rho\sqrt{s}\delta + \delta^2}{s_{\max}(\Sigma^*_{SS}) - \Delta} + \frac{1-\rho}{p-s} - \\ & - \max_{\lambda \in T^{p-s-1}} \left|\lambda^\top(\Sigma_{S^cS^c} - \Sigma^*_{S^cS^c})\lambda\right|. \end{aligned} \tag{1.168}$$

In order to control Δ (1.159), we apply (1.156) with the choices

$$z = \sqrt{s \vee \log n}, \quad \text{and} \quad t = 1 \vee \frac{\log n}{s}.$$

Consequently, there exists a constant $C_1 > 0$ depending only on σ so that if $n > C_1(s \vee \log n)$, we have that

$$\mathbf{P}(\mathcal{A}) \geq 1 - 3\exp(-c'(s \vee \log n)),$$

$$\text{where } \mathcal{A} = \left\{\Delta \leq 2\sqrt{\frac{s^2(1 + 1 \vee (\log(n)/s))}{n}} + C'\sqrt{\frac{s \vee \log n}{n}}\right\} \tag{1.169}$$

In order to control δ (1.160) and the last term in (1.168), we make use of (1.151), which yields that

$$\mathbf{P}(\mathcal{B}) \geq 1 - \frac{6}{p \vee n}, \quad \text{where}$$
$$\mathcal{B} = \left\{ \delta \leq C\sqrt{\frac{s\log(p \vee n)}{n}} \right\} \cap \left\{ \sup_{\lambda \in T^{p-s-1}} \left| \lambda^\top (\Sigma_{S^c S^c} - \Sigma^*_{S^c S^c}) \lambda \right| \leq C\sqrt{\frac{\log(p \vee n)}{n}} \right\}. \tag{1.170}$$

For the remainder of the proof, we work conditional on the two events $\mathcal{A}$ and $\mathcal{B}$. In view of (1.169) and (1.170), we first note that there exists $C_2 > 0$ depending only on σ and ρ such that if $n \geq C_2 s^2 \log(p \vee n)$ the two conditions (1.166) and (1.167) supposed to be fulfilled previously indeed hold. To conclude the proof, we re-write (1.168) as

$$\tau^2(S) \geq \frac{\rho(1 - \Delta/(1-\rho))}{(1 - \Delta/(1-\rho)) + s\frac{\rho}{1-\rho}} + \frac{2\rho\frac{\sqrt{s}}{1+(s-1)\rho}\delta}{1 - \Delta/(1+(s-1)\rho)} - \frac{\delta^2/(1+(s-1)\rho)}{1 - \Delta/(1+(s-1)\rho)} - \\ - \max_{\lambda \in T^{p-s-1}} \left| \lambda^\top (\Sigma_{S^c S^c} - \Sigma^*_{S^c S^c}) \lambda \right|. \tag{1.171}$$

Conditional on $\mathcal{A} \cap \mathcal{B}$, there exists $C_3 > 0$ depending only on σ and ρ such that if $n \geq C_3(s^2 \vee (s \log n))$, when inserting the resulting scalings separately for each summand in (1.171), we have that

$$\begin{aligned} &c_1 s^{-1} - C_4\sqrt{\frac{\log(p \vee n)}{n}} - C_5\frac{\log(p \vee n)}{n} - C_6\sqrt{\frac{\log(p \vee n)}{n}} \\ &= c_1 s^{-1} - C_7\sqrt{\frac{\log(p \vee n)}{n}}. \end{aligned} \tag{1.172}$$

We conclude that if $n \geq \max\{C_1, C_2, C_3\} s^2 \log(p \vee n)$, (1.172) holds with probability no less than $1 - \frac{6}{p \vee n} - 3\exp(-c'(s \vee \log n))$. □

1.4.8 Empirical performance on synthetic data

We here present the results of numerical experiments with synthetic data which are intended to illustrate our analysis of the performance of NNLS in terms of prediction, estimation and sparse recovery. Specific attention is paid to the performance relative to the non-negative lasso. In particular, we provide scenarios in which the latter is visibly outperformed by NNLS. Differences arise mainly with regard to estimation in ℓ_∞-norm and support recovery as already indicated by the discussion in §1.4.5. It is also demonstrated that thresholded NNLS in combination with an entirely data-driven choice of the threshold constitutes a reliable tuning-free procedure for support recovery.

Deconvolution of spike trains. We consider a positive spike train deconvolution model as it commonly appears in various fields of applications, one of which is discussed in depth in §1.5. For the sake of better illustration, we here confine ourselves to

synthetic data and defer a discussion of various additional issues arising in real world datasets. As starting point one considers an underlying signal f which 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 $\{\beta_k^*\}_{k=1}^s$ 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. 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.

$$y_i = \sum_{k=1}^{s} \beta_k^* \phi_k(u_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.173}$$

and if X has the self-regularizing property, it follows immediately from (1.51) in Theorem 1.21 that the ℓ_2-prediction error of NNLS is bounded as

$$\frac{1}{n}\|\mathsf{f} - X\widehat{\beta}\|_2^2 \leq \mathcal{E}^0 + C\sqrt{\frac{\log p}{n}}, \qquad \text{where } \{\mathsf{f}_i = f(u_i)\}_{i=1}^n, \tag{1.174}$$

where $\mathcal{E}^0 = \min_{\beta \geq 0} \frac{1}{n}\|\mathsf{f} - X\beta\|_2^2$. When concluding (1.174) from (1.51), we have absorbed $\|\beta^0\|_1$ into C. This makes sense from the following considerations about the asymptotic regime of interest. One is given a fixed signal which is sampled at a higher rate as n increases, i.e. the resolution becomes finer. In order to improve the localization of the $\{\mu_k\}_{k=1}^s$, it is necessary to increase the number of candidate positions $\{m_j\}_{j=1}^p$. For example, one may take the $\{m_j\}_{j=1}^p$ as the sampling points, in which case $p = n$. 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 $\mathcal{E}^0$ being negligible as long as the $\{m_j\}_{j=1}^p$ are placed densely enough. The bound (1.174) then implies that $\widehat{\beta}$ must have large components only for those columns of X corresponding to locations near the $\{\mu_k\}_{k=1}^s$, which can then be estimated even more accurately by applying a simple form of post-processing as discussed in the next section. By contrast, with ordinary least squares one cannot take $p = \Theta(n)$ candidate positions as n increases and hope for an error bound like (1.174). Hence, accurate localizations of $\{\mu_k\}_{k=1}^s$ based on NNLS can potentially be done with far less sampling points than would be required by ordinary least squares. On the other hand, the application of fast rate bounds such as that of Theorem 1.24 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, restricted eigenvalue of Condition 1.23. For our simulation study, we consider model (1.173). 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

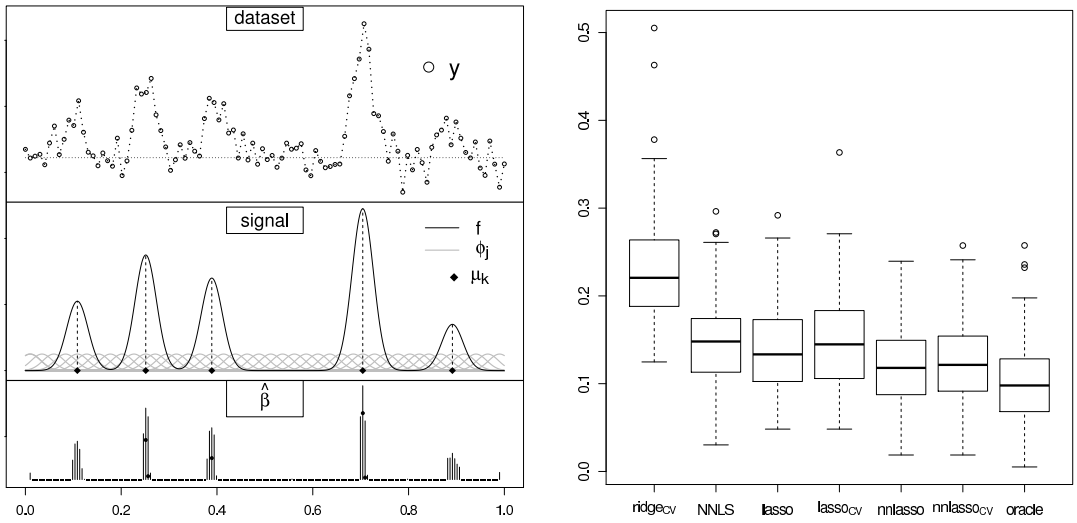


Figure 1.6: Left panel: Visualization of the experimental setting. The middle part depicts the underlying signal, a positive combination of five Gaussians. The upper part depicts a sample dataset generated according to model (1.173). The lower part provides a summary of the coefficient vectors $\widehat{\beta}$ 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).

by construction that $\{m_j\}_{j=1}^p \supset \{\mu_k\}_{k=1}^s$ so that $\mathcal{E}^0 = 0$. 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 (1.46), 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 1.6 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 Figure 1.6 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.

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: Ens_+
The matrix X is generated by drawing its entries independently from the uniform distribution on $[0, \sqrt{3}]$ such that the population Gram matrix has equi-correlation structure with $\rho = 3/4$. Random matrices of that form have been considered for previous numerical results (cf. Figures 1.3 to 1.5). For given (n, p, s), the target β^* is generated by setting $\beta_j^* = 6b \cdot \phi_{\min}^{-1/2} \sqrt{2\log(p)/n}(1 + U_j)$, $j = 1, \dots, s$, where $\phi_{\min} = (1 - \rho)$ denotes the smallest eigenvalue of the population Gram matrix, the $\{U_j\}_{j=1}^s$ are drawn uniformly from $[0, 1]$, and we let the parameter $b > 0$ vary. We further set $\beta_j^* = 0$, $j = (s + 1), \dots, p$.
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 above to fit into the standard sparse recovery framework. Indeed, in the preceding experiments, 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 [24] 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 for the preceding experiment, we consider sampling points $u_i = i/n$, $i = 1, \dots, n$, in $[0, 1]$ and localized functions $\phi_j = \phi(\cdot - m_j)$,where here $\phi(\cdot - m_j) = \exp(-|\cdot - m_j|/h)$, $j = 1, \dots, p$ with $h = 2/n$. The centers m_j, $j = 1, \dots, p$, 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 $m_1, \dots, m_s$ corresponding to S are drawn from the uniform distributions on these intervals. The remaining centers $m_{s+1}, \dots, m_p$ 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 $X_{ij} = \phi_j(u_i)/c_j$, $i = 1, \dots, n$, $j = 1, \dots, p$, 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 $\beta_j^* = b \cdot \beta_{\min}(1 + U_j)$, $j = 1, \dots, s$ and $\beta_j^* = 0$, $j = s+1, \dots, p$. 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, cf. end of §1.1.4) 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 1.1 and 1.2) and ℓ_2-norm (Tables 1.3 and 1.4). 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 1.34 (with σ replaced by

the naive estimator $\frac{1}{n}\|y - X\widehat{\beta}\|_2^2$) 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 e.g. in [11] 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 [55] 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 1.36, 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.

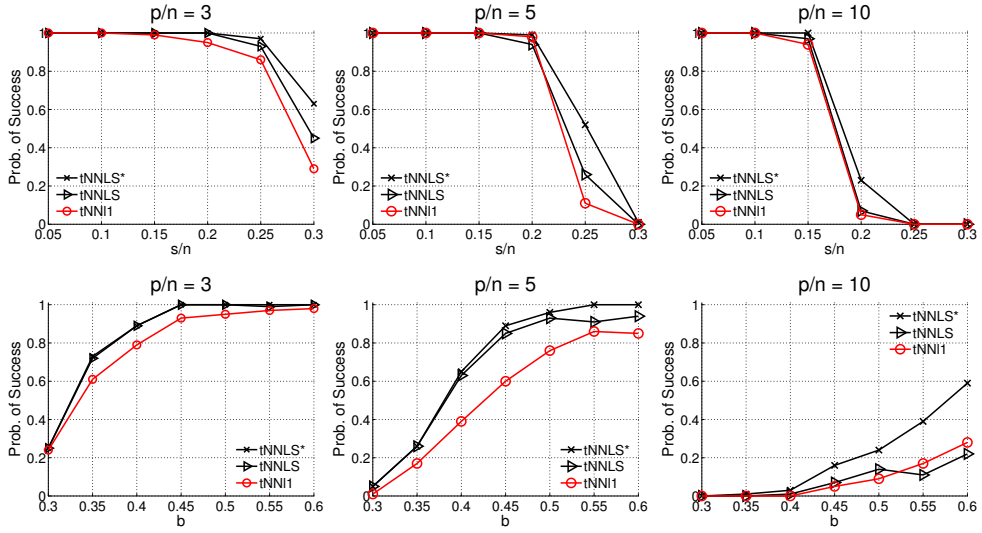


Figure 1.7: Sparse recovery results for Design I (Ens$_+$). Top: Results of the set of experiments with fixed signal strength $b = 0.5$. Bottom: Results of the set of experiments with fixed fraction of sparsity $s/n = 0.2$. 'tNNLS*' and 'tNNℓ_1'denote thresholded NNLS and the thresholded non-negative 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.

Discussion of the results. In summary, Figures 1.7 and 1.8 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 charac-

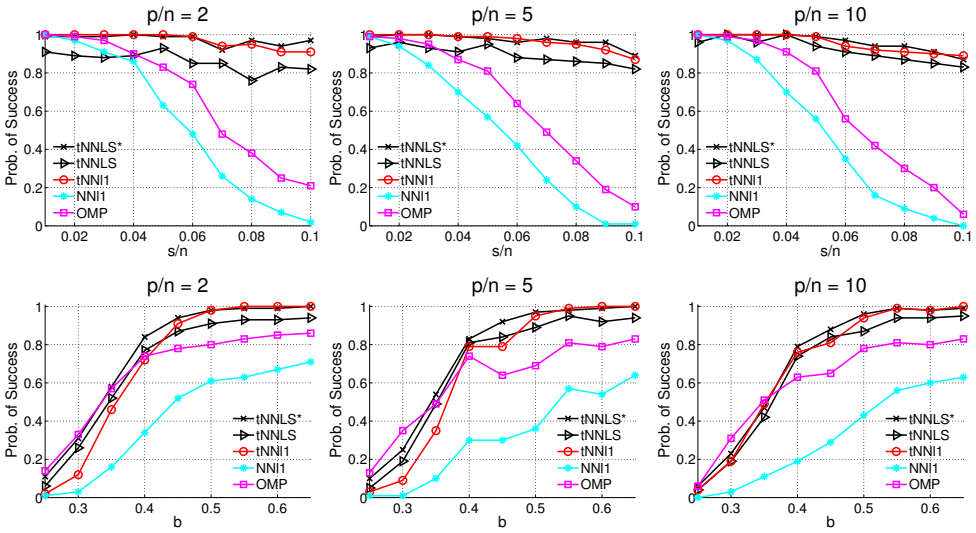


Figure 1.8: Sparse recovery results for Design II (localized non-negative functions). Top: Results of the set of experiments with fixed signal strength $b = 0.55$. Bottom: Results of the set of experiments with fixed fraction of sparsity $s/n = 0.05$. 'tNNLS*' and 'tNNℓ_1'denote thresholded NNLS and 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.

terized by comparatively small signal strength b or high 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 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 [110, 154, 176, 181, 184]. 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 of 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. Proposition 1.35), which is necessary for the success of OMP as well [175], fails to hold in all these runs. The ℓ_∞-errors in estimating β^* reported in Tables 1.1 and 1.2 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 in Tables 1.3 and 1.4.

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

Table 1.1: Averages (± standard errors) of $\|\widehat{\beta} - \beta^*\|_\infty$ (NNLS) and $\|\widehat{\beta}^{\ell_1^+,\lambda} - \beta^*\|_\infty$ (NNℓ_1) for Design I (Ens$_+$) with $b = 0.5$.

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

Table 1.2: Averages (± standard errors) of $\|\widehat{\beta} - \beta^*\|_\infty$ (NNLS) and $\|\widehat{\beta}^{\ell_1^+,\lambda} - \beta^*\|_\infty$ (NNℓ_1) for Design II (localized non-neg. functions) with $b = 0.55$.

	p/n							
	2		3		5		10	
s/n	nnls	nnℓ_1	nnls	nnℓ_1	nnls	nnℓ_1	nnls	nnℓ_1
0.05	1.0±.01	1.0±.01	1.1±.01	1.1±.01	1.2±.01	1.2±.01	1.3±.01	1.3±.01
0.1	1.4±.01	1.4±.01	1.6±.01	1.6±.01	1.8±.02	1.8±.02	2.1±.02	2.1±.02
0.15	1.8±.01	1.8±.02	2.0±.02	2.0±.02	2.4±.02	2.4±.02	3.1±.04	3.4±.05
0.2	2.1±.02	2.2±.04	2.5±.02	2.6±.07	3.1±.03	3.3±.04	5.4±.10	6.9±.19
0.25	2.5±.02	2.6±.04	3.1±.03	3.7±.14	4.5±.07	7.2±.27	12.0±.2	15.3±.2
0.3	3.0±.03	3.4±.11	4.0±.05	5.5±.24	8.1±.19	12.8±.3	18.6±.1	19.8±.1

Table 1.3: Averages (± standard errors) of $\|\widehat{\beta} - \beta^*\|_2$ (NNLS) and $\|\widehat{\beta}^{\ell_1^+,\lambda} - \beta^*\|_2$ (NNℓ_1) for Design I (Ens$_+$) with $b = 0.5$.

	p/n							
	2		3		5		10	
s/n	nnls	nnℓ_1	nnls	nnℓ_1	nnls	nnℓ_1	nnls	nnℓ_1
0.02	0.6±.01	0.7±.01	0.6±.01	0.7±.01	0.6±.01	0.7±.01	0.6±.01	0.7±.01
0.04	0.7±.01	1.0±.01	0.7±.01	1.0±.01	0.7±.01	1.0±.01	0.7±.01	1.0±.01
0.06	0.8±.01	1.2±.01	0.8±.01	1.2±.01	0.8±.01	1.2±.02	0.9±.01	1.2±.01
0.08	0.9±.01	1.3±.02	0.9±.01	1.3±.02	0.9±.01	1.3±.02	1.0±.01	1.4±.01
0.1	1.0±.01	1.4±.02	1.0±.01	1.5±.02	1.0±.01	1.5±.02	1.1±.01	1.5±.02

Table 1.4: Averages (± standard errors) of $\|\widehat{\beta} - \beta^*\|_2$ (NNLS) and $\|\widehat{\beta}^{\ell_1^+,\lambda} - \beta^*\|_2$ (NNℓ_1) for Design II (localized non-neg. functions) with $b = 0.55$.

1.4.9 Extensions

In light of promising theoretical properties and empirical success, it is worthwhile to explore possible extensions of NNLS. Below, we collect a few ideas that either build directly upon NNLS or that make use of concepts such as non-negativity, sparsity and self-regularization. One of these extensions has meanwhile been published [143] while the others are left as topics of future research.

Least squares regression with half-space constraints. The constraint set of NNLS can be represented as an intersection of half-spaces: we have $\mathbb{R}^p_+ = \{x \in \mathbb{R}^p : \cap_{j=1}^p \langle e_j, x\rangle \geq 0\}$. More generally, one can consider constraint sets of the form $\{x \in \mathbb{R}^p_+ : \cap_{j=1}^q \langle a_j, x\rangle \geq 0\}$ for given $a_j \in \mathbb{R}^p$, $j = 1, \ldots, q$. This prompts the following generalization of NNLS:

$$\min_{\beta:\, A\beta \geq 0} \|y - X\beta\|_2^2, \tag{1.175}$$

where $A \in \mathbb{R}^{q\times p}$ contains the $\{a_j\}_{j=1}^q$ as its rows. Accordingly, statistical analysis involves the sparsity of $A\beta^*$ in place of the sparsity of β^*. As an example, let A represent the first order difference operator, i.e.

$$A = \begin{pmatrix} -1 & 1 & 0 & \ldots & \ldots & 0 \\ 0 & -1 & 1 & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & -1 & 1 \end{pmatrix}$$

Then the constraint set is given by all β satisfying $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_p$, while sparsity means that the sequence $\{\beta_j\}_{j=1}^p$ has few 'jumps', i.e. $\beta_j = \beta_{j+1}$ for most j. The central question to be answered is whether – under suitable conditions on X and A – minimizers of (1.175) enjoy similar performance guarantees as NNLS.

In the sequel, we consider three extensions involving convex cones of real matrices.

Non-negative least squares approximations for matrices. Let $Y \in \mathbb{R}^{m_1\times m_2}$ and $Z_j \in \mathbb{R}^{m_1\times m_2}$, $j = 1, \ldots, p$, be given matrices and suppose that we wish to find a non-negative combination of the $\{Z_j\}_{j=1}^p$ optimally approximating Y in a least squares sense. This yields the optimization problem

$$\min_{\beta\geq 0} \left\| Y - \sum_{j=1}^p \beta_j Z_j \right\|_F^2, \tag{1.176}$$

where $\|M\|_F = (\sum_{a,b} M_{ab}^2)^{1/2}$ denotes the Frobenius norm of matrix M. Actually, (1.176) is not a proper extension of NNLS as it can be converted to a standard NNLS problem (1.16) by vectorizing Y, vectorizing the $\{Z_j\}_{j=1}^p$ and stacking them as columns of a corresponding design matrix. Nevertheless, it is worth mentioning (1.176) because of its usefulness for solving several projection problems on polyhedral cones of real

matrices considered in the literature [111, 126] such as the set of diagonally dominant matrices with positive diagonal, or interesting subsets thereof like the set of Laplacian matrices

$$\mathcal{L}^m = \left\{ A \in \mathbb{R}^{m\times m} : \; A = A^\top,\, a_{ii} = -\sum_{j\neq i} a_{ij},\; i = 1,\dots,m,\;\; a_{ij} \leq 0 \text{ for all } i \neq j \right\} \tag{1.177}$$

It is not hard to verify that the set $\mathcal{L}^m$ can equivalently be expressed as

$$\mathcal{L}^m = \left\{ A \in \mathbb{R}^{m\times m} : \; A = \sum_{k=1}^{m}\sum_{l>k}^{m} \lambda_{kl}(e_k e_k^\top + e_l e_l^\top - e_k e_l^\top - e_l e_k^\top), \lambda_{kl} \geq 0 \,\forall k,l \right\}. \tag{1.178}$$

Accordingly, the Euclidean projection of $Y \in \mathbb{R}^{m\times m}$ on $\mathcal{L}^m$ can be recast as a problem of the form (1.176) with

$$\begin{aligned} &Z_1 = e_1 e_1^\top + e_2 e_2^\top - e_1 e_2^\top - e_2 e_1^\top, \quad Z_2 = e_1 e_1^\top + e_3 e_3^\top - e_1 e_3^\top - e_3 e_1^\top, \dots, \\ &Z_p = e_{m-1} e_{m-1}^\top + e_m e_m^\top - e_{m-1} e_m^\top - e_m e_{m-1}^\top, \end{aligned}$$

where $p = m(m-1)/2$. Using the conic hull representation (1.178) may be advantageous over approaches which try to compute the projection based on the half-space representation (1.177), because in the former case it is possible to make use of fast solvers available for NNLS.

Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random fields. Besides regression, sparsity has become a key concept for various other statistical estimation problems. Estimation of the covariance matrix or its inverse (also known as the precision matrix) of a random vector is a traditional task in multivariate analysis that poses a considerable challenge in the high-dimensional setting. Let $\mathbf{Z} = (Z_j)_{j=1}^p \sim N(\mu^*, \Sigma^*)$ be a p-dimensional Gaussian random vector and suppose that we want to estimate its precision matrix $\Omega^* = (\omega^*_{jk}) = (\Sigma^*)^{-1}$ from samples $\{z_1, \dots, z_n\}$ which are i.i.d. realizations of $\mathbf{Z}$, assuming that μ^* is known. Then maximum likelihood estimation can be shown to be equivalent to the log-determinant divergence minimization problem

$$\min_{\Omega \in \mathbb{S}_+^p} -\log\det(\Omega) + \operatorname{tr}(\Omega S), \quad S := \frac{1}{n}\sum_{i=1}^{n}(z_i - \mu^*)(z_i - \mu^*)^\top, \tag{1.179}$$

where $\mathbb{S}_+^p$ denotes the set of by $p \times p$ symmetric positive definite matrices, cf. [74], §17.3. Once $n < p$, the sample covariance matrix S is singular, and as a result, (1.179) is unbounded from below so that a maximum likelihood estimator fails to exist. Moreover, in this situation one cannot hope to reasonably estimate Ω^* unless it possesses additional structure that can be exploited. Sparsity of the off-diagonal entries of Ω^* has primarily been considered in this context, and different forms of sparsity-promoting regularization have been proposed [26, 59, 131]. In the Gaussian setting, sparsity of the off-diagonal entries of Ω^* has a particularly convenient interpretation in terms of the induced conditional independence graph in which pairs of variables (Z_j, Z_k), $k \neq j$, are

connected by an edge if and only if Z_j and Z_k are conditionally independent given the remaining variables $\{Z_l\}_{l\notin\{j,k\}}$, which can be shown to be equivalent to $\omega^*_{jk} = 0$, cf. [90], §5. Thus, sparsity of Ω^* translates to sparsity of the associated conditional independence graph. Independent of Gaussianity, one can also show that $(-\omega^*_{jk}/\omega^*_{jj})_{k\neq j}$ equals the vector of regression coefficients of the linear regression in which Z_j is regressed on $\{Z_k\}_{k\neq j}$, $j = 1,\dots,p$. In our recent paper [143], we specialize to the case in which all these regression coefficients are non-negative. Equivalently, we consider the following subset of $\mathbb{S}^p_+$:

$$\mathcal{M}^p = \{\Omega = (\omega_{jk}) \in \mathbb{S}^p_+ :\ \omega_{jk} \leq 0,\ \ j,k = 1,\dots,p,\ \ j \neq k\},$$

the set of symmetric positive definite M-matrices [8]. In [143], we investigate whether the sign constraints on the off-diagonal elements can be exploited in estimation, and whether adaptation to sparsity similar as in non-negative regression is possible. Specifically, as a direct modification of (1.179), we consider sign-constrained log-determinant divergence minimization

$$\min_{\Omega\in\mathcal{M}^p} -\log\det(\Omega) + \operatorname{tr}(\Omega S). \tag{1.180}$$

We show that, under a mild condition on the sample covariance matrix S, there exists a unique minimizer of (1.180) even if $n < p$. Moreover, we provide theoretical and empirical evidence indicating that thresholding of the off-diagonal entries of the resulting minimizer may be a suitable approach for recovering the sparsity pattern of an underlying sparse target Ω^*.

Recovering symmetric positive definite matrices of low rank. The field of compressed sensing started with the problem of recovering a sparse vector from incomplete linear measurements, but was readily extended to the more general problem of recovering a low rank matrix $B^* \in \mathbb{R}^{m_1\times m_2}$ from linear measurements of the form $y_i = \operatorname{tr}(X_i B^*)$ for certain measurement matrices $X_i \in \mathbb{R}^{m_2\times m_1}$, $i = 1,\dots,n$, cf. e.g. [29, 115, 127]. While sparsity now refers to the singular values of B^*, symmetric positive definiteness appears to be the natural counterpart to non-negativity in the vector case. First recovery results into this direction have been shown in [166], but these fall behind those established for trace norm regularization (the counterpart to non-negative ℓ_1-regularization) in [25, 37]. Besides, the results in [166] are limited to a noiseless setting.

1.5. Sparse recovery for peptide mass spectrometry data

This section is devoted to an in-depth discussion of a real world application of NNLS to a challenging deconvolution problem in proteomics. While theory developed in earlier sections helps to understand the empirical success of the proposed approach, the present section clearly reveals that in order to obtain a reliable system in practice, additional steps need to be performed. While these steps can in principle be regarded as low-level details of the approach, they have an important influence on the performance. In a broad sense, this issue can be attributed to the fact that assumptions typically made in theory, notably correctness of the model, are not exactly met in practice.

1.5.1 Background

Mass spectrometry (MS) is a key technology in the biomedical sciences for analyzing the protein content of biological samples. As such, it plays an important role in systems biology and clinical research, where it is used, among other things, to discover biomarkers and to enhance the understanding of complex diseases. This however only constitutes the last out of a series of steps that are involved with sample preparation, the measurement process, the recording and the analysis of the resulting spectra. A central step in the pre-processing of MS data all subsequent analyses depend on is the detection of the biologically relevant components within the raw data generated by the spectrometer. This task essentially amounts to separating 'signal' from 'noise' and is typically referred to as *feature extraction*. The latter term is rather descriptive since in fact the vast majority of the recorded data points can be classified as noise. In this section, we are concerned with *peptide spectra*. Peptides are chains of amino acids, the building block of proteins. During sample preparation, proteins are divided into multiple peptides via an enzymatic reaction. The thus resulting peptides are separated by a laboratory technique called liquid chromatography, which considerably simplifies subsequent data analysis by reducing the number of overlapping signals to a minimum extent. As made precise below, such overlapping signals are difficult to deal with in feature extraction. For more details on the measurement process, we refer to [139].

1.5.2 Challenges in data analysis

The data sets (spectra) under consideration in this section are composed of intensities observed for a large number of mass-per-charge (m/z) positions, which is typically in the ten to the hundred thousands. The feature selection problem is to detect those m/z-positions at which a peptide is located and to assign charge states (z) resulting from ionization. In combination, one obtains a list of peptide masses, which constitutes a first step in determining the protein-related contents. The signal triggered by peptides becomes manifest in the form of isotopic patterns: the chemical elements serving as building blocks of peptides naturally occur as isotopes differing in the number of neutrons and hence (approximately) by an integer of atomic mass units, such that a peptide produces a signal at multiple mass positions, which becomes manifest in a series of regularly spaced peaks. The task to be performed in data processing is illustrated in Figure 1.9, which displays an excerpt of a MALDI-TOF[10] spectrum recorded from a sample of myoglobin. Upon visual inspection of the raw data (top panel), one identifies five to six clusters of peaks clearly standing out of noise regions. The bottom panel depicts the systematic decomposition of the given excerpt according to the constituent isotopic patterns, as returned by our method. It is shown that the task of properly unmixing the signal into its constituents is more intricate than it might have appeared at first glance. While the three rightmost isotopic patterns are well-separated and thus easy to identify, the pattern around position 970 turns out to be a superposition of two distinct patterns. The occurrence of such overlapping signals makes the problem highly non-trivial as naive approaches that focus on locating peak clusters are not sufficient.

[10]MALDI is a common ionization technique in mass spectrometry. TOF (time of flight) refers to the type of mass spectrometer that is used in conjunction with MALDI.

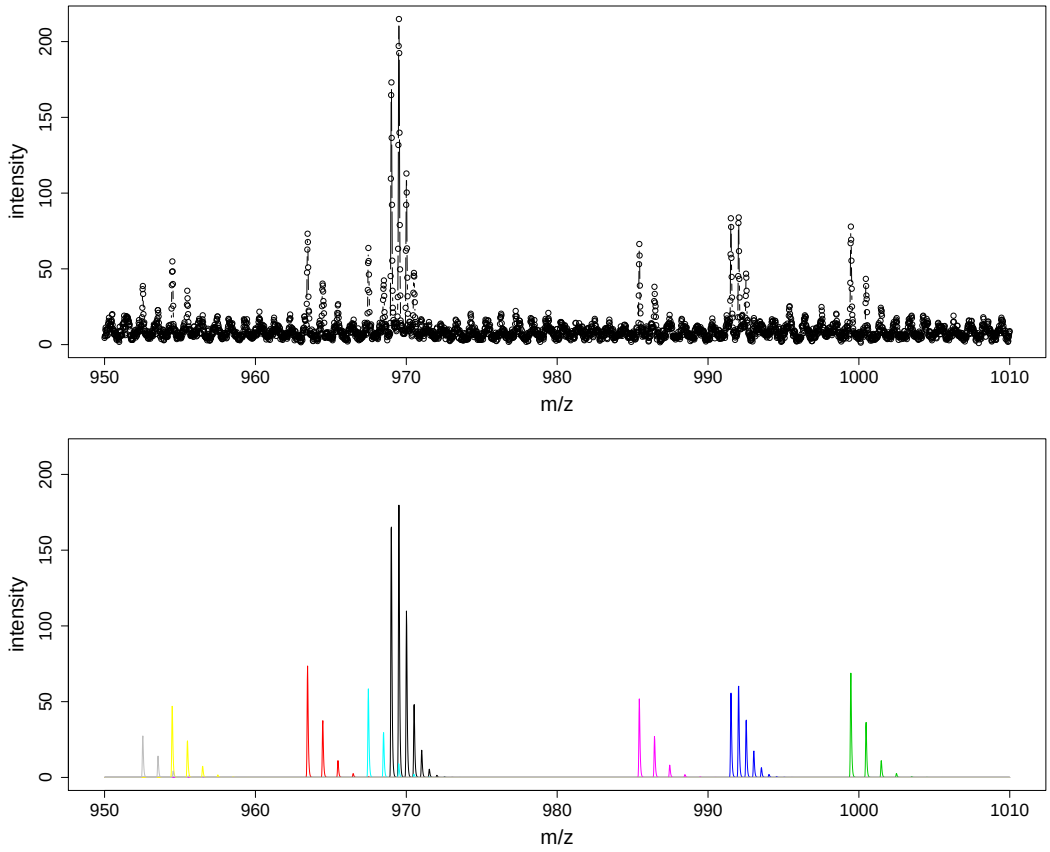


Figure 1.9: Excerpt of a raw spectrum (top) and its decomposition into isotopic patterns (bottom) as returned by our method. Each colour corresponds to a distinct isotopic pattern.

In addition to being able to resolve overlapping signals, an appropriate method for the problem under consideration should have the following properties.

Insensitivity to noise. In the excerpt depicted in Figure 1.9, the signal-to-noise ratio is comparatively large so that it is easy to discriminate between regions of signals and regions of noise. However, most spectra also contain regions with small signal-to-noise ratios or regions containing spurious peaks or peak series which cannot be associated with an underlying isotopic pattern.

Insensitivity to heteroscedasticity. Absolute intensities tend to vary considerably within a single spectrum. A drastic example is shown in the left panel of Figure 1.10. The shown excerpt contains two isotopic patterns whose absolute signal strengths differ roughly by a factor of 50. Fluctuations in the intensity levels go along with heteroscedasticity of the noise. That is, the noise level tends to increase with the local intensity level. This is illustrated in the middle and the right panel of Figure 1.10. Here, the noise level in one part of the spectrum (middle panel) exceeds the signal level in the other part (right panel). Since it is important not to miss signal of low absolute intensity, feature extraction should be based on a suitable local signal-to-noise ratio.

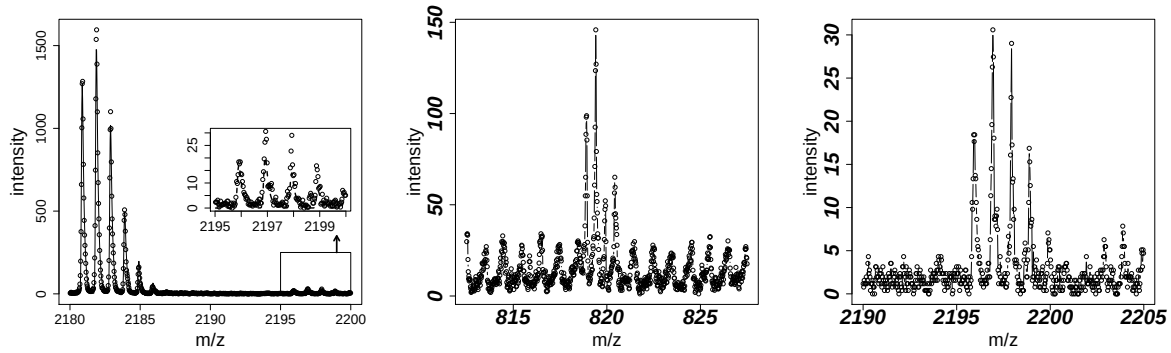


Figure 1.10: Illustrations of heteroscedasticity in peptide mass spectra. Left: Two isotopic patterns in close neighbourhood whose intensities differ drastically. Middle/Right: Two isotopic patterns occurring in different m/z-regions of the same spectrum. Note the different scalings of the vertical axis.

Insensitivity under low resolution. The excerpt of Figure 1.9 has been taken from a high-resolution spectrum. As a result, each peak is represented by quite a few data points. For lowly resolved spectra, the number of data points per peak can be far less, which may substantially complicate signal detection.

User-friendliness. Ultimately, the method should give rise to a software tool to be used by practitioners in proteomics, which do not want to bother about algorithmic details. In particular, application of such tool should not involve laboursome fine-tuning of parameters. Moreover, it is desirable that the tool is insensitive to the specific platform used for data recording.

1.5.3 Formulation as sparse recovery problem

The task of feature extraction as outlined above can be recast as a deconvolution problem. In a continuous domain, the underlying signal composed of s isotopic patterns can be represented as

$$f(x) = \sum_{k=1}^{s} b_k (\psi \star \iota)(x - \mu_k^*), \quad \iota(x - \mu_k^*) := \sum_{l \in \mathbb{Z}} \alpha_l(\mu_k^*; z_k)\, \delta\left(x - \mu_k^* - \frac{l}{z_k}\right), \quad (1.181)$$

where x takes values within some specific interval of m/z-values, the $\{b_k\}_{k=1}^s$ are positive weights (amplitudes) and ψ is a point spread function (PSF) modeling a smeared peak (the default being a Gaussian), which is convolved (denoted by '$\star$') with the function ι. The latter represents an isotopic pattern which is modeled as a positive combination of Diracs centered at m/z-positions $\{\mu_k^* + \frac{l}{z_k}\}$, where the weights $\{\alpha_l(\mu_k^*; z_k)\}_{l \in \mathbb{Z}}$ follow a well-established model for isotopic abundances [142] given the position μ_k^* of the leading peak (i.e. $\alpha_0(\mu_k^*; z_k) \geq \alpha_l(\mu_k^*; z_k), l \neq 0$) and the charge z_k. Depending on the mass $\mu_k^* \cdot z_k$, the sequence $\alpha_l(\mu_k^*; z_k)$ decays quite rapidly with $|l|$, $k = 1, \ldots, s$, so

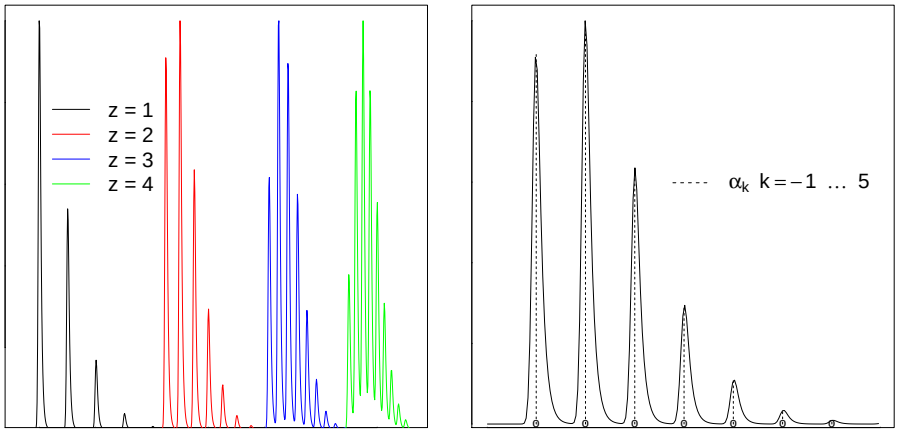


Figure 1.11: Visualization of the model used for isotopic patterns. The left panel depicts isotopic patterns with common m/z-position, but varying charge states z (1 to 4). The right panel provides a more detailed view of the isotopic pattern with $z = 2$.

that isotopic patterns effectively emerge as peak clusters typically consisting of a small number of visible peaks; see Figure 1.11 for a visualization. Sampling of (1.181) at m/z-positions $\{x_i\}_{i=1}^n$ yields a sampled signal $\mathsf{f} = (f(x_1), \dots, f(x_n))$. Note that by sampling, the information about the positions $\{\mu_k^*\}_{k=1}^s$ gets lost. The observed intensities $y = (y_i)_{i=1}^n$ at the sampling points $\{x_i\}_{i=1}^n$ are then given by $y = \mathsf{f} + \varepsilon$, where ε is an error term associated with the measurement process. In terms of model (1.181), the task to be performed is to find the positions $\{\mu_k^*\}_{k=1}^s$ and the corresponding charges $\{z_k\}_{k=1}^s$ as well as the amplitudes $\{b_k\}_{k=1}^s$. For 'benign' spectra, (approximate) deconvolution can be achieved easily in two steps. First, one detects all peaks in y of a significantly high amplitude. Second, nearby peaks are merged to form groups, each group representing an isotopic pattern. The charges $\{z_k\}$ can be inferred from the spacings of the peaks within the same group. For more complicated spectra, this approach is little suitable. Whenever several isotopic patterns overlap, peaks are likely to be overlooked in the first step because of the PSF ψ smearing the peaks out. But even if that does not happen, one cannot hope to correctly assemble detected peaks according to the pattern they belong to in the second step, since nearby peaks may belong to different patterns. Approaches based on *template matching* (see Figure 1.12 for an illustration) circumvent these evident shortcomings by directly tackling the problem at the level of isotopic patterns. In essence, template matching involves a sparse regression scheme in which the design consists of templates matching the shape of isotopic patterns, exploiting that the amplitudes $\{\alpha_l\}$ of the peak series within an isotopic pattern are known given location and charge. Since the composition of the spectrum is unknown in advance, templates are placed at positions $\{\mu_j\}_{j=1}^{\underline{p}}$ covering the whole m/z-range, where $\underline{p}$ is in the order of the number of sampling points n. Since the charge states are unknown as well, one template is used per possible charge state (typically $z \in \{1, 2, 3, 4\}$) at each selected position. This yields the following approximation to the signal:

$$f(x) \approx \sum_{z=1}^{Z} \sum_{j=1}^{\underline{p}} \beta_{z,j}^0 \phi_{z,j}(x), \quad \beta_{z,j}^0 \geq 0, \; z = 1, \dots, Z, \; j = 1, \dots, \underline{p},$$

$$\text{where } \phi_{z,j}(x) = \sum_{l\in\mathbb{Z}} \alpha_l(z;\mu_j)(\psi \star \delta)\left(x - \mu_j + \frac{l}{z}\right),\ z = 1,\dots,Z,\ j = 1,\dots,\underline{p}.$$

Specializing to the sampling points $\{x_i\}_{i=1}^n$, we have

$$\mathsf{f} \approx X\beta^0 = \sum_{z=1}^{Z} X_z \beta_z^0, \quad \text{where } X_z = \big(\phi_{z,j}(x_i)\big)_{\substack{i=1,\dots,n,\\ j=1,\dots,\underline{p}}},\ z = 1,\dots,Z. \tag{1.182}$$

Note that X has $p = \underline{p}\cdot Z$ columns. If $\{\mu_k^*\}_{k=1}^s \subseteq \{\mu_j\}_{j=1}^{\underline{p}}$, the approximation would be exact. The coefficient vector β^0 can then be related to the amplitudes $\{b_k\}_{k=1}^s$ in the sense that $\beta_{z,j}^0 = b_k$ if $\mu_j = \mu_k^*$ and $z = z_k$. Since the number of templates used exceeds by far the number of isotopic patterns present in a spectrum, β^0 is sparse. The term 'template matching' describes the process of identifying those elements of $\{\phi_{z,j}\}$ which have a match in the given signal.

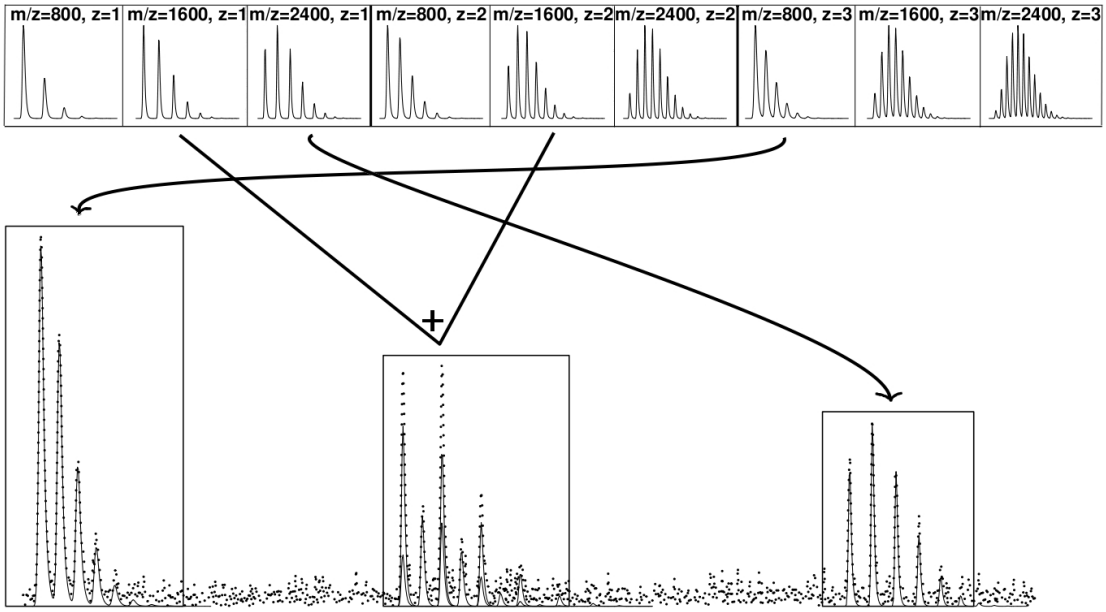


Figure 1.12: Illustration of template matching. The boxes in the top part of the figure contain nine templates $\{\phi_{z,j}\}$ whose shape varies in dependency of mass-over-charge (m/z) and charge (z). The bottom part of the Figure depicts a toy spectrum generated by combining four different templates and adding a small amount of random noise. The arrows indicate how the templates are matched to their counterparts in the spectrum. The signal in the middle is an overlap of two patterns which are accordingly fitted by a combination of templates, which is indicated by a '+'.

If the model were exact, perfect deconvolution could be achieved from the solution of ℓ_0-minimization

$$\min_{\beta\in\mathbb{R}_+^p} \|\beta\|_0 \quad \text{such that } X\beta = \mathsf{f},$$

under mild conditions on X (cf. Proposition 1.1). In the presence of model misspecification and noise, ℓ_0-minimization needs to be replaced by ℓ_0-regularized least squares

$$\min_{\beta \in \mathbb{R}^p_+} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_0.$$

Feature extraction via the non-negative lasso. For the reason of computational intractability, ℓ_0-regularized least squares is not practical. The standard approach is to resort to convex relaxation, i.e. to work with the non-negative lasso problem

$$\min_{\beta \in \mathbb{R}^p_+} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda \mathbf{1}^\top \beta. \tag{1.183}$$

One more rationale behind the use of the non-negative lasso is the idea that the regularizer may provide an effective way of dealing with noise in mass spectra. As can be seen from Figures 1.9 and 1.10, noise tends to arise in the form of some kind of baseline and, because of non-negativity, deviates from the zero-mean assumption usually made in theory. As a consequence, the templates contained in the design matrix can be used to fit noise. The hope is that once the regularization parameter λ is large enough, fitting of noise is prevented and the solution is reasonably sparse. Accordingly, as proposed in [128], one may take the templates corresponding to the non-zero entries of the lasso estimator $\widehat{\beta}^{\ell_1^+,\lambda}$ as the set of extracted features. However, in light of the heteroscedasticity issue discussed above, the non-negative lasso cannot be expected to achieve satisfactory performance. If λ is chosen comparatively large, low-intensity signals are likely to be lost. On the other hand, if λ is chosen such that all low-intensity signals are retained, one has to expect a considerable number of false positive selections which correspond to noise instead of signal. In [128], this problem is attacked by partitioning a given spectrum into bins and solving a non-negative lasso problem separately for each bin. While this strategy solves the issue to some extent, it poses new problems arising from the division of the spectrum. A more direct approach is to choose the amount of regularization locally instead of globally. More specifically, we suggest using a weighted form of ℓ_1-regularization, where the weights are taken according to a local measure of noise. In place of (1.183), we thus consider

$$\min_{\beta \in \mathbb{R}^p_+} \frac{1}{n}\|y - X\beta\|_2^2 + \lambda \sum_{z=1}^{Z} \sum_{j=1}^{\underline{p}} \widehat{\sigma}_j \beta_{z,j}, \tag{1.184}$$

where $\widehat{\sigma}_j$ is a proxy of the 'local noise level' at position μ_j, $j = 1, \ldots, \underline{p}$, which can be obtained e.g. as the median of the intensities within a sliding window (see §1.5.4 for details). To demonstrate that this adjustment is an appropriate way of dealing with heteroscedasticity, we present below the results of a small experiment with synthetic data. For this purpose, we generate data according to the model

$$y_i = 2\phi_1(x_i) + \phi_2(x_i) + 0.5\phi_3(x_i) + \sigma(x_i)\varepsilon_i, \tag{1.185}$$

where the sampling points $\{x_i\}_{i=1}^n$, $n = 5000$, are placed evenly along the m/z-range $[1000, 1150]$. The functions $\{\phi_j\}_{j=1}^3$ represent isotopic patterns of charge $z = 1$ placed

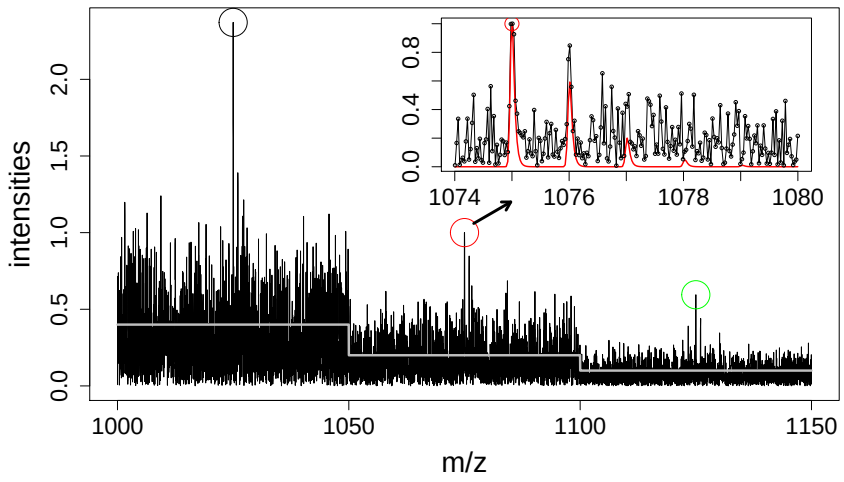


Figure 1.13: An artificial mass spectrum generated randomly according to (1.185). The coloured circles indicate the positions of the initial peak of the patterns (ϕ_1 = black, ϕ_2 = red, ϕ_3 = green). The function σ is drawn in grey.

at the m/z-positions $\{1025, 1075, 1125\}$. The random variables $\{\varepsilon_i\}_{i=1}^n$ are drawn i.i.d. from a zero-truncated Gaussian distribution with standard deviation 0.2. Heteroscedasticity is induced by the positive function σ which is constant on the subintervals $[1000, 1050)$, $[1050, 1100)$, $[1100, 1150]$. Figure 1.13 displays one instance of such a spectrum. The aim is to identify $\{\phi_j\}_{j=1}^3$ from a collection of 600 templates placed evenly in the range $[1000, 1150]$, that is to find the support of β^0 after re-writing (1.185) as

$$y = X\beta^0 + \xi = [X_1\, X_2\, X_3\ \ X_4 \dots X_{600}]\ [\ 2\ \ 1\ \ 0.5\ \ 0\ \ \dots\ \ 0\ ]^\top + \xi,$$
$$\text{where}\ \ y = (y_i),\ \ X_j = (\phi_j(x_i)),\ j = 1, \dots, p,\ \ \xi = (\sigma(x_i)\varepsilon_i).$$

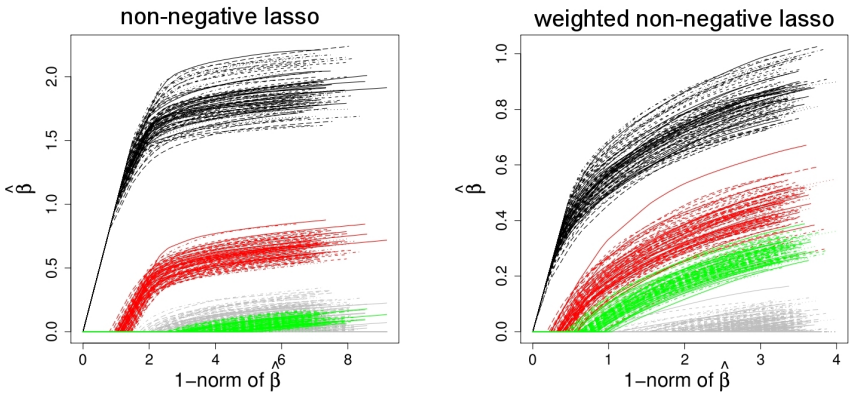


Figure 1.14: Solution paths of the non-negative lasso (left), solution paths of the weighted non-negative lasso (1.184) (right). Colours: ϕ_1 = black, ϕ_2 = red, ϕ_3 = green, off-support templates $\phi_4, \dots, \phi_{600}$= grey.

By construction, ϕ_j is centered at the j-th sub-interval on which σ is constant, $j = 1, \ldots, 3$, while the amplitudes $\{2, 1, 0.5\}$ have been chosen such that the corresponding signal-to-noise ratios are equal. We generate 100 random spectra from (1.185). For each instance, we compute the solution paths [55] of both the non-negative lasso and its weighted counterpart (1.184). For simplicity the $\{\widehat{\sigma}_j\}$ are obtained by evaluating the function σ. The results of the experiments displayed in Figure 1.14 clearly show that ϕ_3 cannot be distinguished from the off-support templates $\phi_4, \ldots, \phi_{600}$ if heteroscedastic noise is ignored. The proposed modification turns out to be an effective means to counteract that problem, since on the right half of the plot, ϕ_3 clearly stands out from the noise.

Thresholded NNLS. In view of a series of positive results in §1.4 regarding the performance of NNLS in high-dimensional regression, it makes sense to perform feature extraction based on NNLS in the following way. In the first step, we compute a NNLS fit to a spectrum, obtaining an estimator $\widehat{\beta}$. For the problem under consideration, the design consists of convex combinations of localized non-negative functions, which combines well with the non-negativity constraints on the coefficient vector (cf. the discussion in §1.4.6 and §1.4.8). However, a NNLS fit alone is not sufficient for feature extraction. As mentioned above, noise can be fitted by the templates contained in the design matrix. Consequently, feature extraction according to the non-zero entries of $\widehat{\beta}$ would yield a vast number of false positive selections. At the same, we expect the coefficients to be small in noise regions. In fact, the non-negativity of both the templates and the coefficients prevent cancellation effects, i.e. the superposition of positive and negative terms to represent values close to zero. In conclusion, hard thresholding of the entries of $\widehat{\beta}$ may be a simple, yet effective strategy to eliminate results of noise fitting in $\widehat{\beta}$. In view of heteroscedasticity, the threshold should be chosen locally. We thus define a thresholded NNLS estimator $\widehat{\beta}(t)$ component-wise by

$$\widehat{\beta}_{z,j}(t) = \begin{cases} \widehat{\beta}_{z,j} & \text{if } \widehat{\beta}_{z,j} \geq t\widehat{\sigma}_j \\ 0 & \text{otherwise,} \quad z = 1, \ldots, Z, \ j = 1, \ldots, \underline{p}, \end{cases} \tag{1.186}$$

where the $\{\widehat{\sigma}_j\}_{j=1}^{\underline{p}}$ represent the local noise levels. This fitting + thresholding procedure offers several advantages over the non-negative lasso.

- One gets around the delicate issue of specifying or tuning the regularization parameter λ. Its choice is intricate because it lacks an intuitive interpretation and is hence hard to grasp for a practitioner. This is unlike the threshold t, which can directly be related to the signal. If the templates are scaled such that $\max_x \phi_{z,j}(x) = 1$ for all z, j, the coefficient $\widehat{\beta}_{z,j}$ equals the estimated amplitude of the highest peak of the template, such that $\widehat{\beta}_{z,j}/\widehat{\sigma}_j$ can be interpreted as signal-to-noise ratio and thresholding amounts to discarding all templates whose signal-to-noise ratio falls below a specific value.
- Feature selection based on the support of the non-negative lasso estimator entails that data fitting and model selection are coupled. While this is often regarded as advantage, since model selection is performed automatically, we think that it

is preferable to have a clear separation between data fitting and model selection, which is a feature of our approach. Prior to thresholding, the output of our fitting approach gives rise to a ranking which we obtain without the necessity to specify any parameter. Selection is completely based on a single fit simply by letting the threshold vary. On the contrary, if one wants to reduce the number of features selected by the lasso, one has to reset the regularization parameter and solves a new optimization problem. Note that it is in general not possible to compute the entire solution path of the lasso [55] for the MS datasets used for the present paper. In fact, the dimension of X is in the ten thousands up to hundred thousands so that the active set algorithm of [55] is prohibitively slow. In this regard, model selection by thresholding is computationally more attractive.

- As already discussed in §1.4.5, the price one has to pay for automatic selection is a non-negligible bias, which complicates the detection of signals with small signal-to-noise ratio. This may constitute a serious issue with respect to the analysis of low-resolution spectra.

- When using the non-negative lasso, rescaling the columns of X is equivalent to solving a weighted non-negative lasso problem. Let $D = \operatorname{diag}(d_1, \dots, d_p)$ be a diagonal matrix containing positive scaling factors $\{d_j\}_{j=1}^p$ on its diagonal and let $X_D = XD$ denote the rescaled design matrix. We then have

$$\min_{\theta \in \mathbb{R}_+^p} \frac{1}{n} \|y - X_D\theta\|_2^2 + \lambda \mathbf{1}^\top \theta = \min_{\beta \in \mathbb{R}_+^p} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \mathbf{1}^\top D^{-1}\beta.$$

  As a consequence, column rescaling has to be done with care, since it affects the amount of regularization used per column. Accordingly, it does matter what kind of column normalization is used. As explained above, normalizing all columns to unit ℓ_∞-norm simplifies the interpretation of the coefficients, but turns out to have a negative effect on the quality of feature extraction. We therefore normalize all columns to unit Euclidean norm. While this is recommended throughout the literature and noticeably improves performance, it cannot be regarded as a canonical choice. In fact, when considering templates of different charge state placed at the same position, their Euclidean norms only depend on their charge states. As a result, normalizing with respect to the Euclidean norm implicitly leads to a preference of certain charge states, which is not desired. By contrast, column rescaling has a trivial effect on NNLS: denoting by $\widehat{\theta}$ the NNLS estimator corresponding to the rescaled matrix X_D, we have $\widehat{\theta} = D^{-1}\widehat{\beta}$.

1.5.4 Practical implementation

In the previous subsection, we have provided a high-level outline of the approach to be taken. Before this approach can be applied in practice, several intermediate steps need to be considered. These steps primarily concern model uncertainty and model misspecification as well as issues related to noise in the data.

Parametric estimation of the PSF. In the previous subsection, we have tacitly assumed that the form of the templates $\{\phi_{z,j}\}$ is precisely known in advance. However,

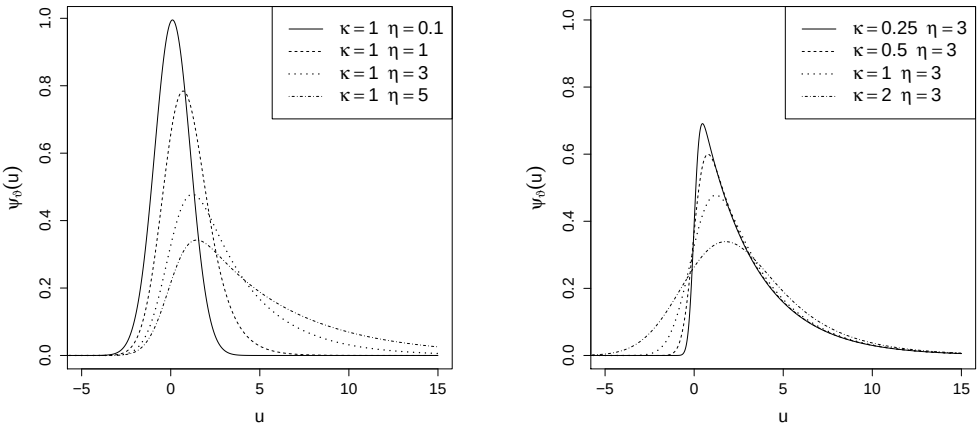


Figure 1.15: Shape of the EMG ($\nu = 0$) for fixed κ and varying η (left panel) and for fixed η and varying κ (right panel).

this is not true in practice. First, the isotope distributions $\{\alpha_l(\mu; z)\}_{l\in\mathbb{Z}}$ (cf. right panel of Figure 1.11) are approximated by an average-case model. Since deviations from that model turn out to be moderate, we stick to this approximation. Second, the PSF ψ is neither known nor is it necessarily the same for all isotopic patterns encountered in the spectrum, i.e. in our model (1.181), it would be more adequate to write $\psi_{\mu_k^*}$ in place of ψ, $k = 1, \ldots, s$. It is common to use a parametric model for the PSF. Among parametric models, a Gaussian PSF is the standard choice. For MALDI-TOF spectra, it is known that the PSF has a pronounced upper tail [185], which does not conform to the symmetry inherent in a Gaussian PSF. We therefore use a more flexible PSF known as exponentially modified Gaussian (EMG), which can be used to model skewed peaks; see for instance [71, 105] and [140] for its use in peptide mass spectrometry. The EMG is parameterized by $\vartheta = (\eta, \kappa, \nu)^\top \in \mathbb{R}_+ \times \mathbb{R}_+ \times \mathbb{R}$ so that

$$\psi_\vartheta(u) = \frac{1}{\eta} \exp\left(\frac{\kappa^2}{2\eta^2} + \frac{\nu - u}{\eta}\right)\left(1 - \Phi\left(\frac{\kappa}{\eta} + \frac{\nu - u}{\kappa}\right)\right), \tag{1.187}$$

where $\Phi(t) = \int_{-\infty}^{t} (2\pi)^{-1/2} \exp(-r^2/2)\, dr$ denotes the cumulative density function of the standard Gaussian distribution. The EMG function is displayed in Figure 1.15 for varying parameter combinations. As can be seen from the figure, the parameter η controls the additional length of the upper tail as compared to a Gaussian. For $\eta \downarrow 0$, the EMG function becomes a Gaussian. The parameter ν is an extra location parameter. The unknown vector of parameters ϑ is allowed to vary with the m/z-position at which isotopic patterns are located in a simple manner, as made precise below. In the sequel, we sketch how to estimate $\vartheta = \vartheta(\mu)$ for a given spectrum. In a first step, we apply a simple peak detection algorithm to the spectrum to identify disjoint regions $\mathcal{R}_r \subset \{1, \ldots, n\}$, $r = 1, \ldots, R$, of well-resolved peaks. For each region, we fit the EMG model (1.187) to the data $\{(x_i, y_i)\}_{i\in\mathcal{R}_r}$ using nonlinear least squares:

$$\min_\vartheta \sum_{i\in\mathcal{R}_r} (y_i - \psi_\vartheta(x_i))^2. \tag{1.188}$$

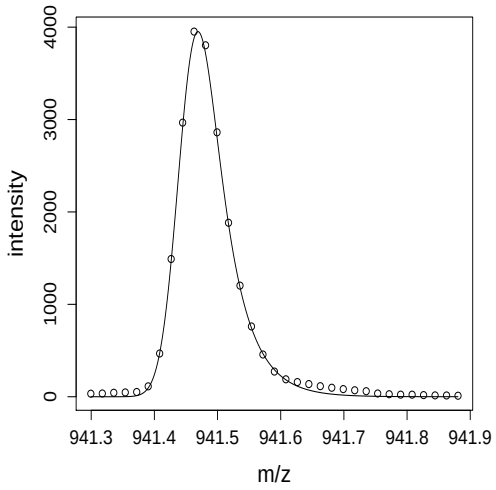


Figure 1.16: Illustration of peak parameter estimation. The figure displays a well-resolved peak in a region consisting of 33 sampling points. The EMG resulting from a nonlinear least squares fit (1.188) is indicated by a solid line.

This is illustrated in Figure 1.16. The solution of (1.188) yields an estimate $\widehat{\vartheta}_r(\widehat{x}_r)$, where $\widehat{x}_r$ denotes an estimate for the mode of the peak in region $\mathcal{R}_r$. Problem (1.188) can be handled with the help of a general purpose nonlinear least squares routine available in popular scientific computing environments. Once all estimates $\{\widehat{\vartheta}_r(\widehat{x}_r)\}_{r=1}^R$ have been obtained, they are subject to a suitable aggregation procedure. In case that the PSF does not depend on the position μ, one could simply take averages. For spectra where peak shape characteristics, in particular peak width, are known to vary systematically with m/z position, we use the pairs $\{(\widehat{x}_r, \widehat{\vartheta}_r(\widehat{x}_r))\}$ as input of a linear regression to infer the parameters of pre-specified trend functions. More specifically, we model each component $\vartheta_l(\mu)$ of $\vartheta(\mu)$ as a linear combination of known functions $g_{l,m}$ of μ and an error component ε_l, i.e.

$$\vartheta_l(\mu) = \sum_{m=1}^{M_l} \gamma_{l,m} g_{l,m}(\mu) + \varepsilon_l(\mu), \tag{1.189}$$

for which a linear trend i.e. $\vartheta_l(\mu) = \gamma_{l,1} + \gamma_{l,2}\mu$, is one of the most common special cases. In [145], a set of instrument-specific models for the peak width is provided, all of which can be fitted by our approach.
We refrain from using least squares regression to determine the parameters in (1.189) due to its sensitivity to possible outliers, which arise from poorly resolved, wiggly or overlapping isotope patterns, which may affect the quality of the initial estimates $\{\widehat{\vartheta}_r\}$. Therefore, the linear model is fitted in a robust way by using least absolute deviation regression [87]. Given the resulting estimates of the parameters $\{\nu_{l,m}\}$, position-specific estimates for the parameters in (1.187) are obtained by evaluating (1.189).

Correcting for effects of model misspecification. As already pointed out above, the linear model used for the signal is not correctly specified. One reason is that the $\{\mu_j\}_{j=1}^p$ which are typically chosen as (a subset of) the sampling points $\{x_i\}_{i=1}^n$ do not contain

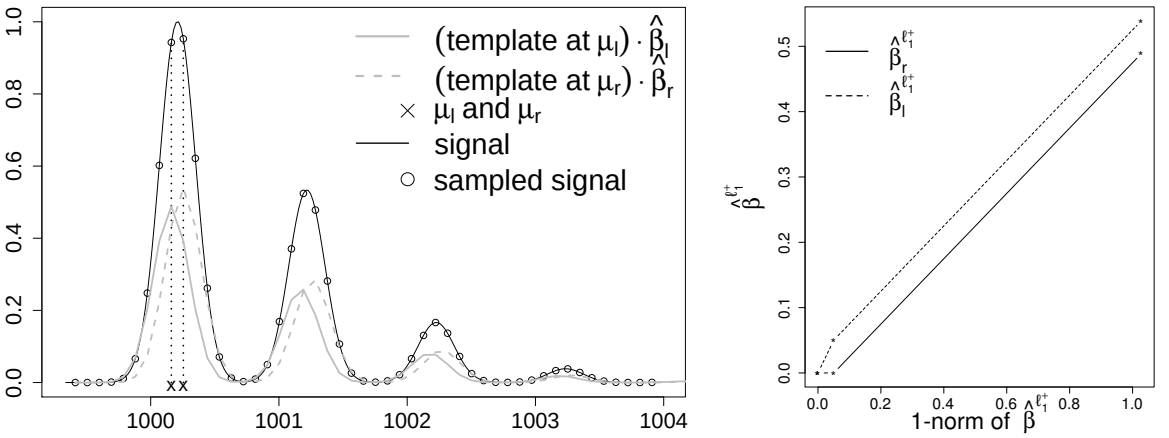


Figure 1.17: Systematic errors in the template model: consequences of a limited sampling rate. The right half of the plot displays the solution path of the non-negative lasso.

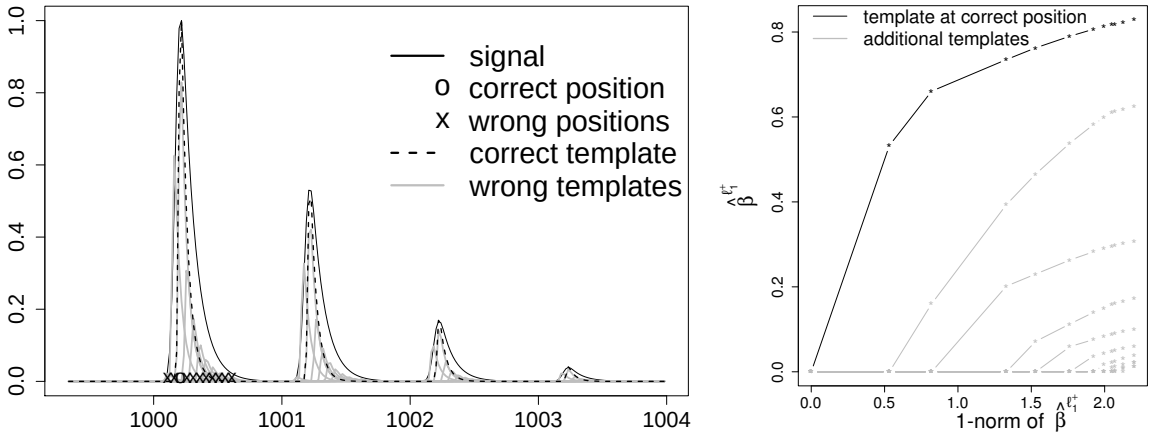


Figure 1.18: Systematic errors in the template model: consequences of an incorrectly specified spread. The right half of the plot displays the solution path of the non-negative lasso.

the $\{\mu_k^*\}_{k=1}^s$. As a result, an approximate model of the form (1.182) can be considerably less sparse because several templates are used to approximate a single isotopic pattern. Accordingly, two or more features would be selected for the same signal. Moreover, this multiplicity with respect to selection implies incorrect quantification of the signal, because its amplitude is distributed among the templates selected. The underlying phenomenon can be understood from the following scenario. Suppose we are given a spectrum corresponding to a single isotopic pattern of amplitude b at position μ^*. Sampling has been performed such that $\mu^* \notin \{x_i\}_{i=1}^n$, and let $\mu_l = \max\{x_i : x_i < \mu^*\}$ and $\mu_r = \min\{x_i : x_i > \mu^*\}$ denote the closest sampling points 'left' respectively 'right' from μ^*. When fitting the sampled signal with two templates placed at μ_l and μ_r using NNLS, the resulting coefficients $\widehat{\beta}_l$ and $\widehat{\beta}_r$ are both assigned positive values depending on the distances $|\mu_l - \mu^*|$ and $|\mu_r - \mu^*|$. In particular, if $|\mu_l - \mu^*| \approx |\mu_r - \mu^*|$ is small, the amplitude b is shared by $\widehat{\beta}_l$ and $\widehat{\beta}_r$ in roughly equal parts. A visualization

is provided in Figure 1.17. That figure also suggests that the non-negative lasso is not an answer to the problem, since only a high amount of regularization leading to a poor fit would achieve a selection of only one template. The second major source of model misspecification results from the fact that the PSF is not known and has to be estimated. Even though the estimation procedure devised in the preceding paragraph yields satisfactory results on average, it may partially induce significant misfits. Figure 1.18 shows the consequences of an underestimation of the spread of the PSF. In order to avoid the effect arising from sampling, we work within an idealized setting where the true m/z-position of the pattern (denoted by 'correct' template in Figure 1.18) is included in the set of positions $\{\mu_j\}_{j=1}^p$ the templates are placed at. Again, ℓ_1-regularization would hardly save the day, because the selection of only one template would underestimate the true amplitude at least by a factor of two, as can be seen from the right panel. We stress that for the situations depicted in Figures 1.17 and 1.18, noise is not present. The issues shown are solely the consequence of model misspecification. The frequent occurrence of the two phenomena described above asks for a correction. One may wonder whether a sparser placement of the templates would mitigate the problem. However, this would affect the accuracy of the approach with regard to the location of the $\{\mu_k^*\}_{k=1}^s$. Moreover, in case of overlapping patterns two nearby templates are in fact required to represent the signal. In light of this, we instead propose a post-processing procedure that aims at undoing the effect of model misspecification by replacing multiple templates fitting a single isotopic pattern by a single template that best approximates the fit in a least squares sense.

Algorithm 1.1 Post-processing

Input: Coefficient vector $\widehat{\theta}$ of an estimation procedure fitting a linear model given (X, y).

$\widehat{S}_z \leftarrow \{j : \widehat{\theta}_{z,j} > 0\}$, $z = 1, \dots, Z$.

for $z = 1, \dots, Z$ **do**

$\overline{\mu}_z \leftarrow 0, \quad \overline{\theta}_z \leftarrow 0.$

Partition $\widehat{S}_z$ into G_z groups $\mathcal{G}_{z,1}, \dots, \mathcal{G}_{z,G_z}$ by merging adjacent positions [11] $\{\mu_j : j \in S_z\}$.

for $m = 1, \dots, G_z$ **do**

Using numerical integration, solve the nonlinear least squares problem

$$(\overline{\mu}_{z,m}, \overline{\theta}_{z,m}) = \operatorname*{argmin}_{\mu,\theta} \left\| \theta \cdot \phi_{z,\mu} - \sum_{l \in \mathcal{G}_m} \widehat{\theta}_{z,l} \phi_{z,l} \right\|_{L_2}^2, \text{[12]} \tag{1.190}$$

where $\phi_{z,\mu}(x) = (\psi_\vartheta \star \iota)(x - \mu)$ is a template at position μ.[13]

end for

end for

return $\{\overline{\mu}_z\}_{z=1}^Z$ and $\{\overline{\theta}_z\}_{z=1}^Z$.

[11] Positions are said to be adjacent if their distance on the m/z scale is below a certain tolerance `ppm` specified in parts-per-million.

[12] $\|f\|_{L^2} := \left(\int_{\mathbb{R}} f(x)^2 \, dx\right)^{1/2}$

[13] Before solving (1.190), the parameters in $\vartheta = \vartheta(\mu)$ can be fixed to one of those templates to be merged; the PSF can be assumed to be the same locally and hence so can be $\vartheta(\mu)$.

In this manner, we not only eliminate multiplicity with respect to selection, but also hope to obtain more accurate estimates of the $\{\mu_k^*\}_{k=1}^s$. As detailed in Algorithm 1.1, all selected templates of the same charge that are within a neighbourhood whose size is proportional to the average spacing of two sampling points are merged to form a group. For each group of templates, precisely one new template is returned that comes closest to the fit when combining all templates of the group, thereby reducing the number of templates returned to only one per detected pattern. By choosing the size of the neighbourhood in the same order of magnitude as the spacing between two sampling points, we ensure that the procedure does not erroneously merge templates that actually belong to different patterns, i.e. no false negatives are introduced at this stage. Furthermore, by taking into account the coefficients of the templates assigned before merging, the accuracy of the position estimates can be considerably improved. Besides, the post-processing procedure can be applied regardless of the specific approach used for data fitting.

Estimation of the local noise level. A suitable measure for the local noise level (henceforth abbreviated as 'LNL') is required for the non-negative lasso with heteroscedasticity adjustment (1.184) as well as for thresholding of the NNLS estimator in (1.186). Moreover, selection of the positions $\{\mu_j\}_{j=1}^p$ at which templates are placed is guided by the LNL. The LNL is taken as the median of the intensities y_i falling into a sliding window of fixed width around a specific position. Formally, given sampling points $\{x_i\}_{i=1}^n$, we set for $x \in [x_1, x_n]$

$$\mathrm{LNL}(x) = \mathrm{median}(\{y_i : \ i \in \mathcal{I}_x\}), \quad \text{where } \mathcal{I}_x = \{i : \ x_i \in [x-h, x+h]\}, \tag{1.191}$$

where the window width $h > 0$ is a parameter. In particular, we use $\widehat{\sigma}_j = \mathrm{LNL}(\mu_j)$, $j = 1, \ldots, p$, in (1.184) and (1.186). In (1.191), the median is preferred over the mean for the reason of robustness. Similar measures for the local noise level can be found in the literature on mass spectrometry, see [85] where a truncated mean is used. Given the LNL, the $\{\mu_j\}_{j=1}^p \subseteq \{x_i\}_{i=1}^n$ are chosen such that $x_i \in \{\mu_j\}_{j=1}^p$ if $y_i \geq \texttt{factor.place} \cdot \mathrm{LNL}(x_i)$, $i = 1, \ldots, n$. That is, templates are placed at position x_i (one for each charge state) if the corresponding y_i exceeds $\mathrm{LNL}(x_i)$ by a factor `factor.place`. In a subsequent paragraph, it is outlined how the two parameters h and `factor.place` can be calibrated in practice. The parameter `factor.place` can in principle be set to zero, in which case $\{\mu_j\}_{j=1}^p = \{x_i\}_{i=1}^n$. However, in order to reduce the computational effort involved in subsequent template matching, it is reasonable to filter out points where one does not expect signal. On the other hand, the choice of the parameter h is more delicate, since an inappropriate choice may have an adverse effect on the quality of feature extraction. Choosing h too small typically has the effect that the signal-to-noise ratio is underestimated such that true peaks might be incorrectly classified as noise. Conversely, choosing h too large typically leads to an overestimation, thereby increasing the number of spurious patterns that are selected. As a rough guideline, h should be chosen proportional to the average envelope size (see (1.194) and the explanation below for a definition) of all templates.

Choice of the loss function. To keep matters simple, we have so far only considered a least squares data fitting term, or squared loss for short. While the use of squared loss is

standard and convenient for computation, it is not clear whether this is a suitable choice from a modelling point of view. It is hard to dismiss the idea that the error distribution is possibly heavy-tailed because of model misspecification that partially causes gross errors. Consequently, it may make sense to use absolute loss in place of squared loss, which is rather sensitive to such gross errors. Accordingly, we consider non-negative least absolute deviation (NNLAD) as an alternative to NNLS. The NNLAD problem is given by

$$\min_{\beta \in \mathbb{R}^p_+} \|y - X\beta\|_1. \tag{1.192}$$

Problem (1.192) can be recast as a linear program by introducing additional variables and can hence be handled by a variety of solvers available for this class of optimization problems. The NNLAD problem compares unfavourably to the NNLS problem from an optimization point of view because the ℓ_1-norm is not smooth; see the last paragraph of this subsection for a more detailed account.

Calibration of parameters. Independent of the specific approach used for template matching (i.e. non-negative lasso, NNLS etc.), we have introduced parameters related to the computation of the local noise level (window width h, cf. (1.191)), template placement (`factor.place`), and the merging scheme used in post-processing (Algorithm 1.1). In order to specify these parameters properly, we perform a grid search, which proceeds as follows. First, one or several spectra are recorded for the purpose of parameter calibration. The resulting data sets are subject to a visual inspection by a human expert who screens the recorded data for isotopic patterns and lists the positions and associated charge states of what he or she classifies as signal. This process is referred to as manual annotation, which constitutes one possible way of generating 'ground truth' for MS data (see §1.5.5 for a discussion). Thereafter, each grid point is evaluated by comparing the set of extracted features in subsequent template matching to the manual annotation, counting true and false positives for a wide range of choices of the threshold t (NNLS) or the regularization parameter λ (non-negative lasso). The corresponding sequence of true/false positive counts is then aggregated into a single score with the help of a summary measure like the fraction of true positives at the precision-recall-break-even-point or the area under the ROC curve, see e.g. [104], §8 for a definition of these terms.

Computational aspects. We now describe our approach for solving the NNLS problem and the corresponding problem in which the squared loss is replaced by the absolute loss. A suitable optimization algorithm needs to scale up to problems of substantial size. The number of sampling points n of a simple spectrum can be of the order of 10^5 and the number of templates p is usually of the same order of magnitude. At the same time, a suitable algorithm should deliver high-accuracy solutions. Low-accuracy solutions tend to be sufficient in regions of strong signal, but may fail to provide accurate estimates of amplitudes in regions of small to moderate signal, which, however, concerns a substantial portion of isotopic patterns. The quality of feature extraction in these regions eventually discriminates between excellent and just average methods. These considerations prompt us to use Newton's method with a log(arithmic) barrier to incorporate the non-negativity constraints. This is a well-established approach, which

belongs to the class of interior point methods ([16], Ch. 11.3). Newton's method in general requires $O(p^3)$ flops per iteration[14], which would be impractical for the given problem size on conventional computers. However, as explained in more detail below, the matrix of templates X is of a sparse structure one can take advantage of. Even though the computational complexity per iteration remains comparatively high relative to other methods using only gradient information (so-called first order methods), Newton's method can achieve considerably more progress per iteration by making use of curvature information. As a result, much less iterations may be required for a high-accuracy solution than for first-order methods.

In the sequel, we first spell out the details of the algorithmic approach for NNLS before discussing the computational complexities of the essential operations involved. A similar scheme is then discussed for absolute loss.

Non-negative least squares. The log barrier method solves a constrained optimization problem with a smooth objective by a reduction to a sequence of unconstrained optimization problems. The main idea is to represent the constraints approximately by a sequence of log barrier terms so that in the limit, these terms become indicator functions of the constraint sets. Here, the indicator functions take the value zero if the argument is contained in the respective constraint set and $+\infty$ otherwise. Specifically, this strategy yields the following sequence of optimization problems for NNLS parameterized by $\gamma > 0$:

$$\min_{\beta \in \mathbb{R}^p} f_\gamma(\beta), \quad \text{where} \quad f_\gamma(\beta) = \frac{1}{2}\|y - X\beta\|_2^2 - \frac{1}{\gamma}\sum_{j=1}^{p} \log(\beta_j),$$

where the second term in f_γ is the log barrier with the convention $\log x = \infty$ for $x \leq 0$. We have rescaled the least squares objective for convenience. In the limit $\gamma \to \infty$, one recovers the NNLS problem. In practice, one starts with a comparatively small value for γ, which is then gradually increased by repeated multiplication with a constant factor larger than one until γ exceeds a predefined upper bound. Newton's method is used to minimize f_γ for each value of γ in the finite sequence defined in this manner. Note that for any $\gamma > 0$, the function f_γ is (strictly) convex on its domain. The gradient and Hessian of f_γ with respect to β, respectively, are given by

$$\nabla f_\gamma(\beta) = -X^\top(y - X\beta) - \frac{1}{\gamma}[1/\beta_1 \ \ldots \ 1/\beta_p]^\top.$$

$$\nabla^2 f_\gamma(\beta) = X^\top X + \frac{1}{\gamma}\mathrm{diag}(1/\beta_1^2, \ldots, 1/\beta_p^2).$$

The Newton direction $d(\beta)$ is obtained from the linear system

$$\nabla^2 f_\gamma(\beta) d(\beta) = -\nabla f_\gamma(\beta). \tag{1.193}$$

Accordingly, the updates are of the form $\beta^{t+1} \leftarrow \beta^t + \alpha^t d(\beta^t)$, where β^t and β^{t+1} denote the iterates at iterations t and $t+1$, respectively, and $\alpha^t > 0$ is a step size determined via an inexact line search based on the Armijo rule ([9], p.29). It is not necessary to obtain highly accurate solutions for intermediate values of γ. It suffices to obtain a

[14]The term flop is an acronym for 'floating point operation', the standard unit used in numerical linear algebra ([69], §1.2.4) to quantify arithmetic operations

final iterate that serves as good starting point for the subsequent value of γ.
Computational complexity. In our implementation, we exploit that the templates contained in the matrix X are highly localized. Therefore, we sparsify X by setting all entries below a certain threshold y_0 (we use $y_0 = 10^{-5}$) equal to zero. Let

$$\begin{aligned} A_{z,j} &= \max\{i : \phi_{z,j}(x_{i-l}) < y_0,\, l = 1, \dots, i-1\}, \\ B_{z,j} &= \min\{i : \phi_{z,j}(x_{i+l}) < y_0,\, l = 1, \dots, n-l\}, \;\; z = 1, \dots, Z,\; j = 1, \dots, \underline{p}, \\ K &= \max_{1 \le z \le Z} \max_{1 \le j \le \underline{p}} B_{z,j} - A_{z,j}. \end{aligned} \tag{1.194}$$

The difference of $B_{z,j}$ and $A_{z,j}$ is called the 'envelope size' of the respective template. The envelopes are sets of consecutive integers $\{A_{z,j}, A_{z,j}+1, \dots, B_{z,j}\} \subseteq \{1, \dots, n\}$ representing the effective support of the templates, cf. the right panel of Figure 1.11. We suppose that the maximum envelope size K is much smaller than the total number of sampling points n. In this situation, substantial savings in both computation and storage can be be achieved by using sparse matrix representations. The matrix X has no more than $K \cdot p$ non-zero entries. Moreover, if $\{\mu_j\}_{j=1}^{\underline{p}} \subseteq \{x_i\}_{i=1}^{n}$, the Gram matrix $X^\top X$ has no more than $3K \cdot Z \cdot p$ non-zero entries and can be computed in $O(pK^2)$ instead of $O(p^2 n)$ flops, which would be prohibitive. In fact, the envelope of any template cannot overlap with more than $3K \cdot Z$ other templates. Treating Z as a constant, this implies that no more than $O(pK)$ non-zero entries in $X^\top X$ need to be computed, each of which amounts to $O(K)$ flops. Likewise, the residual $y - X\beta$ can be computed in $O(nK)$ and the gradient $\nabla f_\gamma(\beta)$ in $O((n+p)K)$ flops. The main effort goes into solving the linear system (1.193) by repeatedly computing Cholesky factorizations of the Hessian $\nabla^2 f_\gamma(\beta)$. The fact that the templates can be reduced to their envelopes implies that after suitable permutation of the rows/columns of $X^\top X$, the Hessian is a band matrix with bandwidth $O(K)$. As a result, obtaining Cholesky factorizations and solving the linear systems can be done in $O(pK^2)$ flops (e.g. [16], p.670). The subsequent inexact line search involves repeated evaluation of the least squares objective by squaring and summing residuals and thus amounts to $O(nK)$ flops.
Non-negative least absolute deviation. Problem (1.192) can be recast as the following linear program.

$$\min_r r^\top \mathbf{1} \quad \text{sb. to} \quad X\beta - y + r \succeq 0, \quad y - X\beta + r \succeq 0, \quad r, \beta \succeq 0.$$

Adding log-barrier terms for all constraints, the objective of the resulting unconstrained convex problem takes the form

$$\widetilde{f}_\gamma(r, \beta) = r^\top \mathbf{1} - \frac{1}{\gamma} \left(\sum_{i=1}^{n} (\log(\xi_i^+) + \log(\xi_i^-) + \log(r_i)) + \sum_{j=1}^{p} \log(\beta_j) \right),$$

where we have used the notational shortcuts (for simplicity, we suppress dependence on r and β)

$$\xi_i^+ = (X\beta)_i - y_i + r_i, \qquad \xi_i^- = y_i - (X\beta)_i + r_i, \;\; i = 1, \dots, n.$$

The gradients w.r.t. r and β, respectively, are given by

$$\nabla_r \widetilde{f}_\gamma(r,\beta) = \mathbf{1} - \frac{1}{\gamma}\left[\frac{1}{(\xi_1^+ + \xi_1^- + r_1)} \;\cdots\; \frac{1}{(\xi_n^+ + \xi_n^- + r_n)}\right]^\top,$$
$$\nabla_\beta \widetilde{f}_\gamma(r,\beta) = -\frac{1}{\gamma}(X^\top([\Xi^+]^{-1} - [\Xi^-]^{-1})\mathbf{1} + [1/\beta_1 \;\ldots\; 1/\beta_p]^\top), \quad \Xi^\pm := \operatorname{diag}(\xi_1^\pm, \ldots, \xi_n^\pm).$$

Introducing $R = \operatorname{diag}(r_1, \ldots, r_n)$ and $B = \operatorname{diag}(\beta_1, \ldots, \beta_p)$, the Hessian is given by the block matrix

$$\nabla^2 \widetilde{f}_\gamma(r,\beta) = \begin{bmatrix} \underbrace{\frac{1}{\gamma}([\Xi^+]^{-2} + [\Xi^-]^{-2} + R^{-2})}_{=:H_{r\,r}} & \underbrace{\frac{1}{\gamma}([\Xi^+]^{-2}X - [\Xi^-]^{-2}X)}_{=:H_{r\,\beta}} \\ \underbrace{\frac{1}{\gamma}(X^\top[\Xi^+]^{-2} - X^\top[\Xi^-]^{-2})}_{=H_{r\,\beta}^\top} & \underbrace{\frac{1}{\gamma}(X^\top([\Xi^+]^{-2} + [\Xi^-]^{-2})X + B^{-2})}_{=:H_{\beta\,\beta}} \end{bmatrix}.$$

Again, it is suppressed that the blocks $H_{r\,r}$ etc. depend on r and β. The linear system for the Newton descent directions reads

$$\begin{bmatrix} H_{r\,r} & H_{r\,\beta} \\ H_{r\,\beta}^\top & H_{\beta\,\beta} \end{bmatrix} \begin{bmatrix} d_r(r,\beta) \\ d_\beta(r,\beta) \end{bmatrix} = - \begin{bmatrix} \nabla_r \widetilde{f}_\gamma(r,\beta) \\ \nabla_\beta \widetilde{f}_\gamma(r,\beta) \end{bmatrix}.$$

Note that $H_{r\,r}$ is diagonal, so it is a cheap operation to resolve for $d_r(r,\beta)$ once $d_\beta(r,\beta)$ is known:

$$d_r(r,\beta) = -(H_{r\,r})^{-1}(H_{r\,\beta} d_\beta(r,\beta) + \nabla_r \widetilde{f}_\gamma(r,\beta)). \tag{1.195}$$

Plugging this into the second block of the linear system, one obtains

$$-H_{r\,\beta}^\top (H_{r\,r})^{-1}(H_{r\,\beta}\, d_\beta(r,\beta) + \nabla_r \widetilde{f}_\gamma(r,\beta)) + H_{\beta\,\beta} d_\beta(r,\beta) = -\nabla_\beta \widetilde{f}_\gamma(r,\beta)$$

which is equivalent to

$$(H_{\beta\beta} - H_{r\,\beta}^\top (H_{r\,r})^{-1} H_{r\,\beta}) d_\beta(r,\beta) = -\nabla_\beta \widetilde{f}_\gamma(r,\beta) + H_{r\,\beta}^\top (H_{r\,r})^{-1} \nabla_r \widetilde{f}_\gamma(r,\beta).$$

In order to solve this linear system in $d_\beta(r,\beta)$, we proceed as for NNLS. Given $d_\beta(r,\beta)$, we resolve for $d_r(r,\beta)$ according to (1.195). In total, the computational complexity is of the same order as above, because the sparsity pattern of the coefficient matrix remains unchanged. For NNLS, re-computation of the Hessian only involves a diagonal update, an operation of negligible computational cost. However, for NNLAD, re-computation of the block $H_{\beta\,\beta}$ involves the matrix multiplication $(X^\top([\Xi^+]^{-2} + [\Xi^-]^{-2})X)$, which amounts to $O(pK^2)$ flops as does the subsequent solution of the linear system via a Cholesky factorization.

(Weighted) non-negative lasso, cf. (1.183). For both loss functions, a combination with a regularizer of the form $\lambda \mathbf{1}^\top W\beta$, where W is a fixed matrix of positive weights, can be treated similarly. In fact, the only change arises for the gradients which involve the additional term $\lambda W\mathbf{1}$.

1.5.5 Performance in practice

In the present subsection, we assess the practical performance of our template matching-based approach in some of its variants with regard to feature extraction from peptide mass spectra.

Datasets. Our evaluation is based on eight different peptide mass spectra generated on two different platforms, MALDI-TOF and ESI [15]. ESI spectra have been recorded with two different mass analyzers referred to as ion trap (IT) and orbitrap (FT). Samples of bovine myoglobin and chicken egg lysozyme at two different concentration levels (10 and 500 fmol [16]) underlie the MALDI-TOF spectra. For the ESI spectra lysozyme at concentrations 250 and 1000fmol has been used. The use of different platforms, mass analyzers and concentration levels is intended to demonstrate the insensitivity of our approach with respect to changes in the data-generating process. ESI spectra are more challenging than MALDI-TOF spectra, because the former contain also quite a few multiply charged signals, whereas for the latter almost all signals are singly charged.

Validation strategy. Validation of feature extraction is notoriously difficult, because a gold standard which is satisfactory from both statistical and biological points of view is missing. In this context, a major problem one has to account for is that spectra frequently contain patterns whose shape is not distinguishable from those of peptides, but which are in fact various artefacts resulting e.g. from impurities during sample preparation and measurement. These artefacts do not constitute biologically relevant information and are, in this sense, 'false positives'. On the other hand, from a statistical perspective which judges a method according to how well it is able to detect specific patterns in a given dataset, a qualification as 'true positive' is justified. With the aim to unify these aspects, we have worked out a dual validation scheme as detailed below.

Comparison with manual annotation. The first part of the validation scheme tries to capture to what extent a method could replace a human expert who annotates the spectra manually upon visual inspection. Automatically generated lists of peptide masses are matched to the manual annotation of the human expert such that an entry of the list is declared 'true positive' whenever there is a corresponding mass in the manual annotation deviating by no more than a certain instrument-specific tolerance. Otherwise, it is declared 'false positive'. As all methods to be compared depend on a parameter ζ (e.g. threshold or regularization parameter) governing, crudely speaking, the trade-off between precision and recall, we explore the performance for a range of reasonable values for ζ, instead of fixing an (arbitrary) value, which we believe to be little meaningful. The results are then visualized as ROC curve, in which each point in the (Recall, Precision)-plane corresponds to a specific choice of the parameter. Formally, we introduce binary variables $\{B_i(\zeta)\}$ for each mass i contained in the list of cardinality $\widehat{L}(\zeta)$ when setting the threshold equal to t, where $B_i(\zeta)$ equals 1 if the mass

[15] Besides MALDI (footnote [10]), ESI is another common ionization technique in mass spectrometry.

[16] fmol = femtomole = 10^{-15} mole. Standard unit of measurement used for the amount of chemical substances.

is matched and 0 otherwise, and denote by L the number of masses of the manual annotation. The true positive rate (recall, R), and the positive predictive value (precision, P) associated with threshold t are then defined by $R(\zeta) = \frac{\sum_i B_i(\zeta)}{L}$, $P(\zeta) = \frac{\sum_i B_i(\zeta)}{\widehat{L}(\zeta)}$. An ROC curve results from a sequence of pairs $\{R(\zeta), P(\zeta)\}$ for varying ζ, cf. Figures 1.19 and 1.20.

Database query. The second part of the validation scheme evaluates the lists in terms of a query to the Mascot search engine [120]. In a nutshell, peptide mass lists are fed into Mascot which then tries to match these masses to peptides belonging to proteins in its database, which covers common proteins such as myoglobin and lysozyme used here. The more peptides from a particular protein are identified, the higher the score returned by Mascot. In this sense, the Mascot score is used as a surrogate for the capability of a feature extraction method to recover an underlying protein, which is of central interest to practitioners. The database query also accounts for a major problem of a manual annotation, namely that peptides yielding weak signals might easily be overlooked, but might be detected by methods designed to extract those weak signals. As for the comparison with the manual annotation, we evaluate several lists corresponding to different choices of the parameter ζ. Instead of an ROC curve, which turned out to be visually unpleasant, we display the statistics of the Mascot output (score, protein sequence coverage and fraction of hits) of two lists per method, namely of those achieving the best score and the best coverage, respectively.

Methods compared. We compare NNLS and NNLAD (1.192) plus thresholding (1.186) to which we here refer as l_2 and l_1, respectively, and the (weighted) non-negative lasso (1.184). For the latter, the columns of the matrix X are normalized to unit Euclidean norm as it is standard in the literature on the lasso. A grid search over 50 values $\{\lambda_k\}_{k=1}^{50}$ for λ is performed, where the construction of the grid follows [62]. Peptide mass lists are obtained from the active sets $A(\lambda_k) = \{z, j : \widehat{\beta}_{z,j}^{\ell_1^+,\lambda_k} > 0\}$, $k = 1, \ldots, 50$. The output of all three methods is refined with the help of the post-processing procedure of Algorithm 1.1.

Our comparison includes three additional methods that do not fall into our framework because they are not based on template matching.

Pepex. Pepex [135] uses non-negative least squares fitting as well, but it is applied to so-called centroided spectra instead of the raw spectra. During centroiding, some form of peak detection is used to extract all peak clusters from a raw spectrum. At the second stage, called de-isotoping, peak clusters are fitted by a design matrix containing isotope distributions (i.e. weights $\{\alpha_l(\mu_k^*; z_k)\}$, cf. (1.181)) as its columns. For pepex, de-isotoping is based on NNLS. The approach avoids explicit modelling of the PSFs and is hence computationally more attractive as the resulting design matrix is much sparser than a design matrix of templates. However, the division into centroiding and de-isotoping may lead to poor performance for low resolution and noisy data, or in the presence of overlapping patterns. In these cases, peak detection is little reliable. In our template-based approach, there is no separation of centroiding and de-isotoping. It performs much better in the aforementioned cases, since it operates directly on the data and is hence less affected if single peaks of a pattern are difficult to detect. This

reasoning is supported by our evaluation as well as that in [128].
Since pepex is limited to detect patterns of charge state one, its performance is only assessed for MALDI-TOF spectra.
Vendor. For the MALDI-TOF spectra, we use software for automatic feature extraction provided by the vendor of the spectrometer.
Isotope Wavelet. As opposed to our template-based approach, this method is not able to handle overlapping signals. On the other hand, it typically shows strong performance in noisy and low intensity regions or on spectra of low resolution [79].

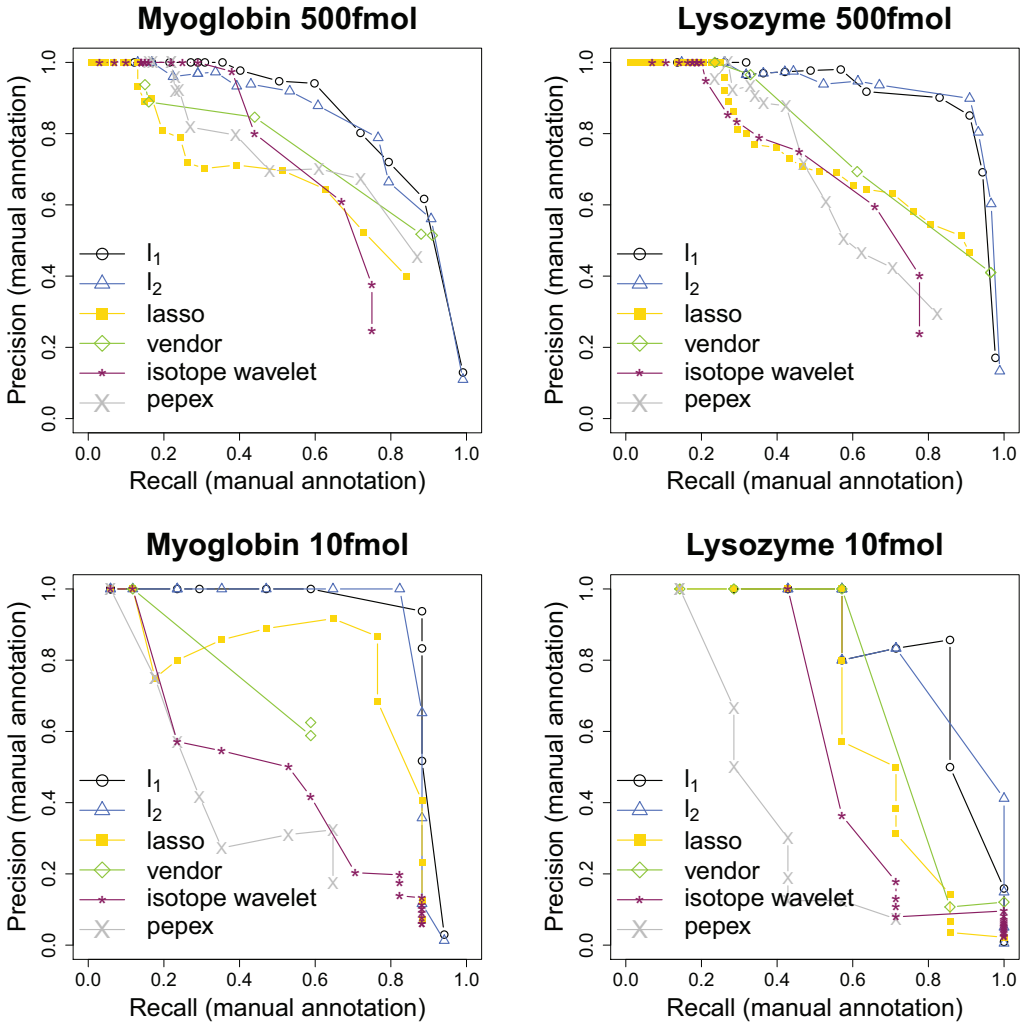


Figure 1.19: Performance for the MALDI-TOF spectra relative to the manual annotation. The points in the (Recall,Precision)-plane correspond to different choices of a parameter ζ controlling the trade-off between precision and recall.

Results. When inspecting Figures 1.19 and 1.20 on the one hand and Table 1.5 on the other hand, one notices that the results of the evaluation based on the manual annotation are not in full accordance with the results of the database query. The

difference is most striking for the MALDI-TOF spectra at 500 fmol, where l_1 and l_2 yield a significant improvement over its competitors, which does not become apparent from the database query. This is because only a fraction of the manual annotation is actually confirmed by the database query. The part which is not matched likely consists of artefacts due to contamination or chemical noise as well as of specific chemical modifications of peptides not present in the Mascot database. In light of this, our dual validation scheme indeed makes sense.

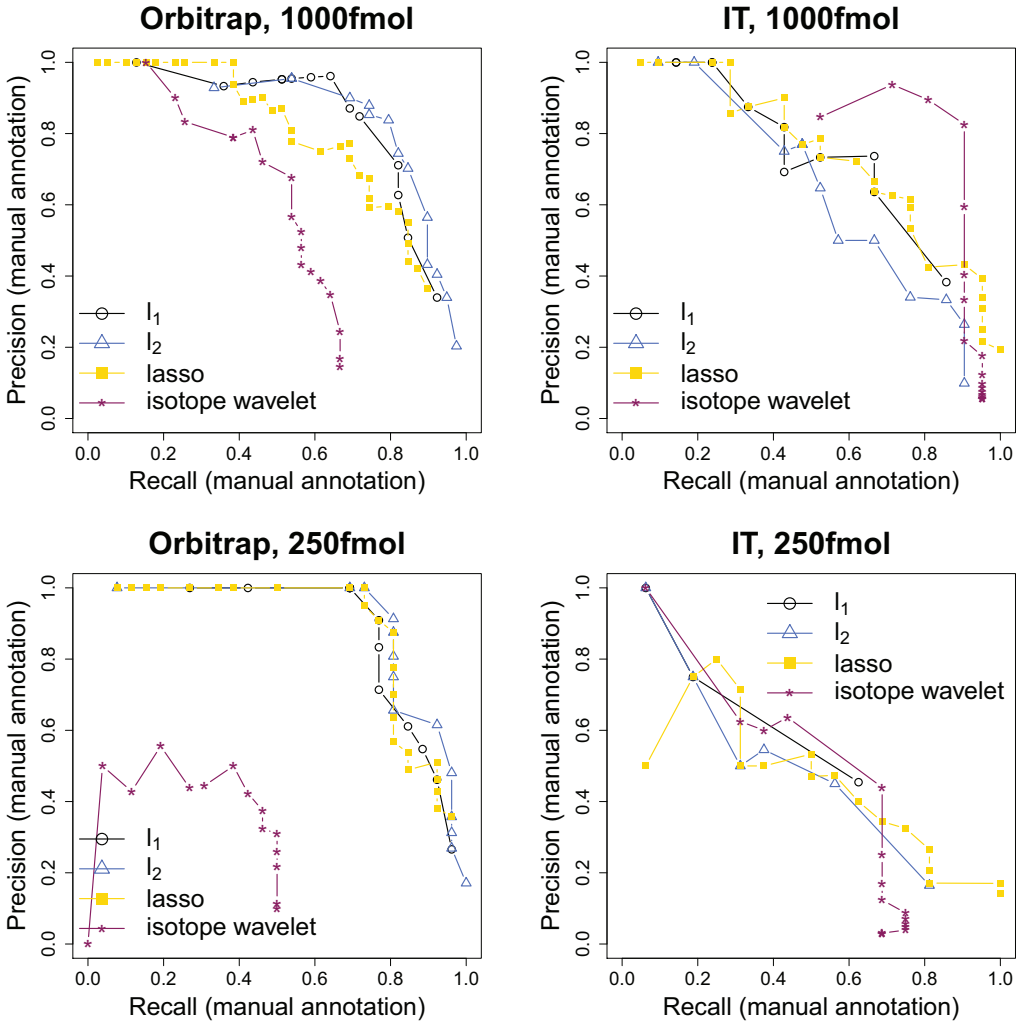


Figure 1.20: Performance for the ESI spectra relative to the manual annotation. The points in the (Recall,Precision)-plane correspond to different choices of a parameter ζ controlling the trade-off between precision and recall.

Comparison. Figure 1.19 and Table 1.5 reveal an excellent performance of our methods l_1 and l_2 throughout all MALDI-TOF spectra under consideration. For the myoglobin spectra a high protein sequence coverage is attained that clearly stands above those of

MALDI Myo 500fmol	score	cvrg	hits	score	cvrg	hits
l_1	211.0	**0.85**	0.94	96.8	**0.96**	0.04
l_2	211.0	**0.85**	0.94	49.6	**0.96**	0.04
lasso	207.0	**0.85**	**1.00**	142.0	0.91	**0.37**
pepex	**223.0**	**0.85**	**1.00**	142.0	0.90	0.17
vendor	**223.0**	**0.85**	0.94	**174.0**	0.90	0.29
wavelet	207.0	**0.85**	**1.00**	156.0	0.90	0.14
MALDI Lys 500fmol	score	cvrg	hits	score	cvrg	hits
l_1	167.0	**0.81**	0.57	133.0	**0.83**	**0.37**
l_2	168.0	0.80	0.64	**144.0**	**0.83**	0.34
lasso	151.0	0.64	0.77	112.0	**0.83**	**0.37**
pepex	**172.0**	0.80	0.63	135.0	**0.83**	0.25
vendor	146.0	0.64	**0.75**	91.4	**0.83**	0.20
wavelet	127.0	0.58	**0.75**	113.0	0.81	0.20
MALDI Myo 10fmol	score	cvrg	hits	score	cvrg	hits
l_1	**211.0**	**0.85**	0.94	82.2	**0.95**	0.04
l_2	207.0	0.74	**1.00**	109.0	0.90	0.14
lasso	195.0	0.77	0.87	**146.0**	0.85	0.46
pepex	97.8	0.80	0.22	97.8	0.80	0.22
vendor	123.0	0.62	0.62	123.0	0.62	**0.62**
wavelet	131.0	**0.85**	0.13	131.0	0.85	0.13
MALDI Lys 10fmol	score	cvrg	hits	score	cvrg	hits
l_1	**89.0**	0.35	**1.00**	**73.7**	0.54	**0.23**
l_2	**89.0**	0.35	**1.00**	35.4	0.72	0.09
lasso	81.9	**0.46**	0.70	46.0	0.74	0.10
pepex	47.1	0.17	**1.00**	31.2	0.53	0.12
vendor	62.7	0.23	**1.00**	43.2	0.34	0.16
wavelet	55.4	0.23	0.45	43.8	**0.82**	0.10
Orbi Lys 1000fmol	score	cvrg	hits	score	cvrg	hits
l_1	149.0	0.70	0.78	138.0	0.80	**0.53**
l_2	139.0	**0.80**	0.50	**139.0**	0.80	0.50
lasso	**159.0**	0.63	**0.87**	120.0	**0.81**	0.29
wavelet	105.0	0.69	0.44	95.1	0.80	0.23
IT Lys 1000fmol	score	cvrg	hits	score	cvrg	hits
l_1	78.7	0.63	0.28	70.9	0.74	0.17
l_2	82.1	0.72	0.36	35.4	0.85	0.13
lasso	103.0	**0.84**	0.33	**76.8**	**0.99**	**0.21**
wavelet	**107.0**	0.79	**0.63**	69.8	**0.99**	0.11
Orbi Lys 250fmol	score	cvrg	hits	score	cvrg	hits
l_1	107.0	0.63	0.50	100.0	0.80	0.31
l_2	103.0	0.63	0.52	66.9	**0.81**	0.14
lasso	**108.0**	0.63	**0.77**	**107.0**	0.80	**0.27**
wavelet	80.6	**0.70**	0.22	80.6	0.70	0.22
IT Lys 250fmol	score	cvrg	hits	score	cvrg	hits
l_1	59.4	0.46	0.16	59.4	0.46	0.16
l_2	37.0	0.59	0.14	37.0	0.59	0.14
lasso	**66.3**	**0.84**	0.20	**66.3**	**0.84**	**0.20**
wavelet	56.3	0.59	**0.36**	21.3	0.75	0.12

Table 1.5: Results of the Mascot database query complementing the comparison to the manual annotation as displayed in Figures 1.19 and 1.20.
The left halves of the tables report the statistics when choosing the parameter ζ to optimize the score, the right halves when optimizing the protein sequence coverage ('cvrg', given as fraction). The column 'hits' contains the fraction of masses that could be matched to peptide masses in the database.

competing methods. For the spectra at 10 fmol, only the performance of the lasso is competitive with that of l_1 and l_2 in terms of the Mascot score; all other competitors, including the vendor software which has been tailored to process these spectra, are significantly weaker. In particular, the strikingly high proportion of 'hits' ($\geq 94\%$) indicates that even at moderate concentration levels, l_1 and l_2 still distinguish well

between signal and noise. This observation is strongly supported by the ROC curves in Figure 1.19, where the precision drops comparatively slowly with increasing recall. In this regard, l_1 and l_2 clearly contrast with the isotope wavelet that aims at achieving high protein sequence coverage. The latter often requires the selection of extremely lowly abundant peptide signals hidden in noise at the expense of reduced specificity.
For MALDI-TOF spectra at high concentration levels, pepex achieves the best scores and is competitive with respect to protein sequence coverage. However, the performance of pepex degrades dramatically at lower concentration levels, as it is unambiguously shown by both parts of the evaluation. In particular, the database scores are the worst among all methods compared.
For the ESI spectra, l_1 and l_2 in total fall a bit short of the lasso (particularly for the ion trap (IT) spectra), but perform convincingly as well, thereby demonstrating that they can deal well with multiple charge states. This is an important finding, since the presence of multiple charges makes the problem much more challenging, because the number of templates as well as the correlations across templates are increased. In spite of these difficulties, Figure 1.20 and Table 1.5 suggest that the performance of the pure fitting approaches l_1 and l_2 does not appear to be affected.

Additional remarks.

- In Figure 1.19, the area under the curve (AUC) of our methods attained for myoglobin is higher for lower concentration. At first glance, this may seem contradictory since an increase in concentration should lead to a simplified problem. However, a direct comparison of the AUCs is problematic, since the number of true positives (17 at 10fmol, 106 at 500fmol) is rather different. For instance, there are choices of the threshold that yield 18 true positives and not a single false positive for both of our methods at 500fmol, yet the AUC is lower.

- The fact that some of the ROCs start in the lower left corner results from outputs containing only false positives.

Chapter 2

Matrix Factorization with Binary Components

Finding an (approximate) factorization of a given data matrix into two factors of low rank is a common way of performing dimension reduction. The (truncated) singular value decomposition (SVD), for example, provides such low rank matrix factorization. The SVD is not suitable once additional constraints like non-negativity are imposed on the two factors. In this case, it is not possible in general to compute the desired factorization. We here study the low rank matrix factorization problem with binary constraints imposed on one of the factors. While at first glance, the combinatorial constraints seem to induce a further complication on top of the nonconvexity of the matrix factorization problem, they turn out to be much more manageable than box constraints, for example. As main contribution, we present an algorithm that provably solves the exact factorization problem in $O(mr2^r + mnr + r^2n)$ flops, where m and n are the dimensions of the input matrix and r is the rank of the factorization. We state conditions for uniqueness of the factorization, in which case the same algorithm can be employed if additional constraints are imposed on the second factor. The question of uniqueness is related to a rich theory revolving around a central result in combinatorics, the Littlewood-Offord lemma.

Chapter outline. We first provide a short account on low-rank matrix factorization which serves as background for the specific problem considered herein. We then present our algorithm for that problem, prove its correctness, analyze its computational complexity and derive conditions under which there is a unique solution. In the sequel, the proposed algorithm is extended to the approximate case and its performance on synthetic data sets is demonstrated. This is complemented by a case study in DNA methylation array analysis.

Notation table for this chapter.

$M \in S^{m\times n}$	M is a matrix with entries M_{ij} from $S \subseteq \mathbb{R}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.
$M_{I,J}$	submatrix of M corresponding to rows in $I \subseteq \{1, \ldots, m\}$ and columns in $J \subseteq \{1, \ldots, n\}$
$M_{I,:}$	submatrix of M corresponding to rows in I
$M_{:,J}$	submatrix of M corresponding to columns in J
$M_{i:i',\, j:j'}$	submatrix of M corresponding to rows $\{i, i+1, \ldots, i'\}$ and columns $\{j, j+1, \ldots, j'\}$
$[M,\, M']$	column-wise concatenation of matrices M, M'
$[M;\, M']$	row-wise concatenation of matrices M, M'
$\mathrm{aff}(M)$	affine hull generated by the columns of M, i.e. $\mathrm{aff}(M) = \{y \in \mathbb{R}^m : y = M\lambda,\ \lambda \in \mathbb{R}^n, \sum_{i=1}^n \lambda_i = 1\}$
$\mathrm{span}(M)$	range of M, i.e. $\mathrm{span}(M) = \{y \in \mathbb{R}^m : y = M\lambda,\ \lambda \in \mathbb{R}^n\}$
$\|M\|_F$	Frobenius norm of M, i.e. $\|M\|_F = \left(\sum_{i,j} M_{ij}^2\right)^{1/2}$
$\|v\|_2$	Euclidean norm of $v \in \mathbb{R}^n$, i.e. $\|v\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$
$\|S\|$	cardinality of a set S
$\mathbb{R}_+$	non-negative real line, i.e. $\{v \in \mathbb{R} : v \geq 0\}$
I_m	$m \times m$ identity matrix
$\mathbf{1}_m$, $\mathbf{1}_{m\times n}$	vector of m ones, $m \times n$ matrix of ones
0_m, $0_{m\times n}$	vector of m zeroes, $m \times n$ matrix of zeroes

Acronyms

HT	HOTTOPIXX
L-O	Littlewood-Offord
NMF	non-negative matrix factorization
RMSE	root mean squared error
SVD	singular value decomposition

2.1. Low-rank representation and the singular value decomposition

Many types of high-dimensional data are intrinsically low-dimensional. For example, images of handwritten digits can be grouped into clusters; facial images consist of a few distinctive parts such as eyes, nose and lips; changes in thousands of gene expression levels can often be explained by changes in a small number of associated functional processes; bag-of-words data from text documents are compactly described via a set of underlying topics; movie ratings depend mostly on few factors, notably genre of the movie and demographic status of the consumer. In times of increasingly complex data sets, low-dimensional representations become indispensable as a means to extract the essential characteristics out of a vast amount of information. Low-dimensional structure typically manifests itself in form of a data matrix that has (approximately) low rank. Suppose we are given $D \in \mathbb{R}^{m\times n}$ and we wish to approximate it by a matrix of rank $r \leq \operatorname{rank}(D) \leq \min\{m,n\}$ with respect to the Frobenius norm, which yields the optimization problem

$$\min_{D' \colon \operatorname{rank}(D') \leq r} \|D - D'\|_F^2. \tag{2.1}$$

Even though the constraint on the rank is non-convex, problem (2.1) can be solved via the singular value decomposition (SVD).

Theorem 2.1. *(SVD, e.g. Theorem 2.5.2. in [69]) Let $M \in \mathbb{R}^{m\times n}$. Then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and a matrix $\Sigma \in \mathbb{R}^{m\times n}$ of the form*

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}})$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min\{m,n\}} \geq 0$, such that

$$M = U\Sigma V^\top.$$

The entries $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ are called the singular values of M. Computation of the SVD is a well-studied problem in numerical linear algebra. Its average runtime complexity is $O(mnr)$, where r is the rank of the input matrix [73]. The next statement establishes an explicit relation between problem (2.1) and the SVD.

Theorem 2.2. *(Eckart-Young Theorem [54]) Consider optimization problem (2.1) and let $D = U\Sigma V^\top$. Then $\widehat{D} = U^{(r)}\Sigma^{(r)}(V^{(r)})^\top$ is a minimizer of (2.1), where $U^{(r)} = U_{:,1:r}$, $\Sigma^{(r)} = \Sigma_{1:r,1:r}$ and $V^{(r)} = V_{:,1:r}$.*

The decomposition $U^{(r)}\Sigma^{(r)}(V^{(r)})^\top$ is sometimes called the r-truncated SVD of D. Note that if $\operatorname{rank}(D) = r$, the approximation is exact, i.e. $\widehat{D} = D$. The truncated SVD can be regarded as a method of linear dimension reduction. Assuming that D has approximately rank r in the sense that its remaining singular values $\sigma_{r+1}, \ldots, \sigma_{\min\{m,n\}}$ are small, we may write

$$D \approx \widehat{D} = U^{(r)} \underbrace{\Sigma^{(r)}(V^{(r)})^\top}_{\Lambda^{(r)}} = U^{(r)}\Lambda^{(r)}, \tag{2.2}$$

i.e. the data points $\{D_{:,1}, \ldots, D_{:,n}\}$ are approximated by a low-dimensional subspace of $\mathbb{R}^m$ spanned by $\{U_{:,1}, \ldots, U_{:,r}\}$. Specifically, we have for $j = 1, \ldots, n$,

$$D_{:,j} \approx \sum_{k=1}^{r} U_{:,k} \Lambda_{kj}.$$

In this context, the $\{U_{:,k}\}_{k=1}^r$ are referred to as factors, components or latent variables. We note that if the data points are centered, i.e. $\sum_{j=1}^n D_{:,j} = 0$, the $\{U_{:,k}\}_{k=1}^r$ coincide with the top r principal components of $\{D_{:,1}, \ldots, D_{:,n}\}$, e.g. [162], Theorem 2.1.

SVD and low-rank matrix factorization. Theorem 2.2 also shows that the low-rank factorization (2.2) provided by the SVD is optimal in the following sense. Consider the problem

$$\min_{T \in \mathbb{R}^{m \times r},\, A \in \mathbb{R}^{r \times n}} \|D - TA\|_F^2. \tag{2.3}$$

Then the pair $(U^{(r)}, \Lambda^{(r)})$ is a minimizer of (2.3). Note that (2.3) is a non-convex problem as is (2.1). In (2.3), non-convexity arises from the fact the objective involves the product of the optimization variables. Nevertheless, it is straightforward to solve (2.3) via the SVD.
In practice, proper use of the truncated SVD remains challenging for at least two reasons. First, the rank r has to be chosen suitably. If r is chosen too large, important features of the original data may get lost. On the other hand, if r is chosen too large the extra components represent negligible details or unstructured information/noise. Second, the Frobenius norm is highly sensitive to outliers.

2.2. Structured low-rank matrix factorization

Principal component analysis (PCA)/truncated SVD was the unrivalled way of performing dimension reduction for years. Advances in the field of optimization have fostered the development of more sophisticated methods, which has been driven by the goal to increase interpretability as well as by the need to have methods of dimension reduction that are somewhat tailored to peculiarities of the underlying data. For example, in the analysis of functional data [121] where each data point represents a smooth function sampled at m points, it is reasonable to enforce smoothness of the principal components. Sparsity of the principal components may be more adequate in other applications, specifically in a high-dimensional setup (cf. Chapter 1), where sparsity needs to be exploited to restore statistical consistency of PCA [81]. If the data points are non-negative, then non-negativity of the components may be beneficial; see the next section for a detailed discussion. The aforementioned properties as well as many others can be incorporated by adding constraints to the low-rank matrix factorization problem (2.3). This yields an optimization problem of the form

$$\min_{T \in \mathcal{C}_T,\, A \in \mathcal{C}_A} \|D - TA\|_F^2, \tag{2.4}$$

where $\mathcal{C}_T \subseteq \mathbb{R}^{m \times r}$ and $\mathcal{C}_A \subseteq \mathbb{R}^{r \times n}$ are constraint sets enforcing or encouraging specific structure such as smoothness, sparsity or non-negativity. Accordingly, we refer to (2.4)

as 'structured matrix factorization problem'. The interpretation of the factors T and A remains straightforward: the columns of T represent the underlying components and the columns of A contain the coefficients of the data points w.r.t. the low-dimensional representation induced by T. On the other hand, because of the additional constraints − even if they are convex − it is not possible in general to compute a global minimizer of (2.4). Consequently, (2.4) has to be treated as an instance of a general nonlinear optimization problem. A common approach is block coordinate descent (cf. Algorithm 2.1 below), where T is optimized for fixed A and vice versa in an alternating fashion. For Algorithm 2.1 to be practical, it is required that it is feasible to evaluate each 'argmin' therein, even though there is a more general variant in which only stationary points instead of minimizers need to be found. In general, little can be said about the quality of the solution returned by Algorithm 2.1, typically not more than convergence to a stationary point of (2.4). In particular, the quality of the solution depends on the initialization. Nevertheless, Algorithm 2.1 is frequently used in practice, notably because of lack of alternatives.

Algorithm 2.1 Block coordinate descent

Initialize T^0

for $t = 0, 1, \ldots$ **do**

$$A^{t+1} \leftarrow \operatorname*{argmin}_{A \in \mathcal{C}_A} \|D - T^t A\|_F^2$$

$$T^{t+1} \leftarrow \operatorname*{argmin}_{T \in \mathcal{C}_T} \|D - T A^{t+1}\|_F^2$$

end for

2.3. Non-negative matrix factorization

Non-negative matrix factorization (NMF) as introduced in [119] is inarguably one of the most popular structured matrix factorization schemes. In its basic form, the NMF problem is given by

$$\min_{T \in \mathbb{R}_+^{m \times r},\, A \in \mathbb{R}_+^{r \times n}} \|D - TA\|_F^2 \tag{2.5}$$

NMF has established itself as the standard method of dimension reduction for non-negative data, i.e. $D \in \mathbb{R}_+^{m \times n}$. In the seminal paper [92], the authors motivate the use of NMF by claiming that it has the property to provide a meaningful 'parts-based decomposition' of the data points. Specifically, the authors observe that for a collection of greyscale facial images, the resulting factors T and A tend to be sparse: the columns of T roughly represent different parts of the face such as eyes, nose and lips in several variations with respect to shape and location; the sparsity in A results from the fact that each single face only contains a subset of all parts relevant to the entire collection. In [92], the authors argue that the observed sparsity is a consequence of the non-negativity imposed on both factors. As a result, cancellation of terms having different signs cannot occur, which automatically promotes sparsity as already discussed in the

first chapter of this thesis. This property turns out to be a key feature in a wide range of applications of NMF; see [38] and [66] for an overview.

Computation. As mentioned above, structured matrix factorization problems pose a computational challenge. For NMF, several hardness results have been established. In [160] the exact NMF problem

$$\text{find } T \in \mathbb{R}_+^{m\times r} \text{ and } A \in \mathbb{R}_+^{r\times n} \text{ such that } D = TA. \tag{2.6}$$

is considered. It is shown that this problem is NP-hard if r is not considered as fixed, hence the existence of an algorithm polynomial in n, m and r is unlikely. More recently, it is shown in [3] that (2.6) can be solved in polynomial time if $r = O(1)$. Moreover, the authors show that if (2.6) could be solved in time $(nm)^{o(r)}$ one could solve 3-SAT in sub-exponential time[1]. In practice, alternating optimization algorithms prevail. When using block coordinate descent (Algorithm 2.1), the block updates for T and A amount to non-negative least squares problems. This scheme is analyzed in [97] where projected gradient is used for the block updates. The multiplicative updates rule of [93] can be rewritten as an alternating gradient descent scheme with a specific choice of the step size which ensures feasibility of the iterates.

Computation under separability. Even though the exact NMF problem (2.6) is computationally hard in general, one may ask whether it becomes easier if there is a solution (T^*, A^*) having additional structure. In [3] it is shown that (2.6) can be solved by linear programming if T^* is *separable*.

Definition 2.3. *A matrix $T \in \mathbb{R}_+^{m\times r}$ is separable if there exists a permutation matrix $\Pi \in \mathbb{R}^{m\times m}$ such that $\Pi T = [\Theta; M]$, where $\Theta \in \mathbb{R}_+^{r\times r}$ is a diagonal matrix with positive entries on its diagonal and $M \in \mathbb{R}_+^{(m-r)\times r}$.*

The notion of separability is originally due to [48] where it is employed in the context of uniqueness of NMF (cf. the following paragraph). Separability is a meaningful assumption for popular applications of NMF. Translated to the facial image collection example above, it means that for each of the constituent parts represented by the columns of T, there is at least one pixel for which the respective part takes a non-zero value, whereas it is zero for the remaining parts. This makes sense if the parts are localized as it is the case in this specific example. To see why separability yields a tremendous simplification from a computational point of view, suppose that $D = T^*A^*$ for T^* separable and observe that

$$\Pi D = (\Pi T^*)A^* = \begin{pmatrix}\Theta\\ M\end{pmatrix} A^* = \begin{pmatrix}\Theta A^*\\ MA^*\end{pmatrix}$$

with Θ and M as in Definition 2.3. We deduce that that − up to positive scaling factors contained in Θ − the rows of A^* appear as the rows of D. Moreover, the remaining rows of D can be expressed as non-negative combinations of the rows of A^* / rows of

[1] 3-SAT is a famous problem in computational complexity theory. It is NP-hard and it is conjectured that its complexity is exponential.

ΘA^*. Consequently, problem (2.6) can be reduced to finding r rows of D whose conic hull contains the remaining rows. One way to perform this task is linear programming: for each row of D, one checks whether it can be expressed as a conic combination of the other rows of D. If this check fails, a row of ΘA^* has been found. This process is iterated until ΘA^* has been determined, and a matching left factor can be found by solving the linear program

$$\text{find } T \in \mathbb{R}_+^{m\times r} \text{ such that } T\Theta A^* = D.$$

While this approach has polynomial time complexity, it requires $O(m)$ linear programs in $O(m)$ variables to be solved, which can be impractical in modern applications. Following the above approach, a single linear program in m^2 variables is suggested in [13] under the name HOTTOPIXX. The authors develop an incremental gradient method with parallel architecture to tackle large-scale problems. The algorithms in [68, 88] try to obtain ΘA^* from the rows of D in a greedy way, which can be significantly faster than linear programming. All these approaches can be extended to the more realistic case where $D \approx T^*A^*$ with T^* (near-)separable. The methods in [3, 13, 68] are backed up with a robustness analysis, and the method in [88] is shown to be robust empirically.

Uniqueness. Given a non-negative matrix factorization TA, one can generate an alternative factorization $(TQ)(Q^{-1}A)$ by choosing Q as a monomial matrix, i.e. $Q = D\Pi$ for a diagonal matrix D having positive diagonal elements and a permutation matrix Π. This does not constitute a major issue because the conic hulls of the columns of T respectively TQ coincide. In this sense, the two factorizations are equivalent. Optional re-scaling can be ruled out by fixing the columns sums of T or the row sums of A. However, if for some non-monomial Q there exists a re-factorization $(TQ)(Q^{-1}A)$ with $TQ \in \mathbb{R}_+^{m\times r}$ and $Q^{-1}A \in \mathbb{R}_+^{r\times n}$, the two factorizations are no longer equivalent. In this case, neither the columns of T nor those of TQ can be interpreted as 'the generating parts'. Having such interpretation is desirable in typical applications of NMF, and is often the central motivation for using it. Unfortunately, it is not clear how to check in practice whether a given NMF is in fact unique (modulo re-scaling and permutation).

2.4. Matrix Factorization with Binary Components

We now turn to the contribution of this chapter. In the sequel, we discuss the problem of computing an approximate factorization of the form

$$D \approx TA, \quad \text{where } T \in \{0,1\}^{m\times r} \text{ and } A \in \mathbb{R}^{r\times n},\ A^\top \mathbf{1}_r = \mathbf{1}_n. \tag{2.7}$$

The additional constraint to have the columns of A sum up to one is imposed for reasons of presentation (see below); it entails that the data points are (approximately) affine instead of linear combinations of the columns of T. In fact, that constraint is not essential to our approach. Motivated by a specific application, the following problem will be addressed as well.

$$D \approx TA, \quad \text{where } T \in \{0,1\}^{m\times r} \text{ and } A \in \mathbb{R}_+^{r\times n},\ A^\top \mathbf{1}_r = \mathbf{1}_n. \tag{2.8}$$

Note that (2.8) is a special instance of an NMF problem. The section title may suggest that we restrict ourselves to the situation where the left factor is binary, but we can likewise deal with a right binary factor, i.e.

$$D \approx TA, \quad \text{where } T \in \mathbb{R}^{m\times r},\ T\mathbf{1}_r = \mathbf{1}_m \text{ and } A \in \{0,1\}^{r\times n}. \tag{2.9}$$

After transposition, we obtain again a factorization with a left binary factor

$$D^\top \approx A^\top T^\top, \quad \text{where } T \in \mathbb{R}^{m\times r},\ T\mathbf{1}_r = \mathbf{1}_m \text{ and } A \in \{0,1\}^{r\times n}. \tag{2.10}$$

The difference between (2.7) and (2.10) is that in (2.7) the data points correspond to the columns of the left hand side, whereas in (2.10) the data points correspond to the rows of the left hand side. The interpretations of (2.7) and (2.10) differ accordingly. In (2.7) each data point is approximated by an affine combination of binary components. On the other hand, (2.10) represents the most basic form of a parts-based decomposition in the sense that parts can be either absent (0) or present (1), but there is no quantification of the 'abundance' of the parts. Both variants are contrasted in Figure 2.1.

2.4.1 Applications and related work

Factorization (2.7) arises in the following applications.

- Our interest in (2.7) originates in a collaboration with researchers in genetics and computational biology regarding the analysis of DNA methylation profiles obtained from high-throughput experiments. Here, each data point represents fractions of DNA methylation (between 0 and 1) at several hundred thousands of sites of the DNA. At a basic level, the DNA methylation profile of a biological sample can be modelled as a mixture of corresponding *binary* methylation profiles of 'pure' cell types, with zero and one indicating absence respectively presence of methylation. These binary matrices form the columns of the matrix T. The columns of A represent the composition (mixture proportions) of the samples w.r.t. the underlying cell types. Since A is non-negative, the desired factorization is actually of the form (2.8).
- A second application is a special case of blind source separation. Here, one is given a set of signals each of which is a superposition of binary source signals. The goal is to recover the sources and to decompose each signal according to these sources. This problem has been discussed in wireless communication [159]. The top panel of Figure 2.1 serves as illustration of such setting.

Factorization (2.9) respectively its equivalent (2.10) arise in the following applications.

- In [141, 155] factorizations of the form (2.9) are considered for the analysis of gene expression data. The idea is to explain changes in gene expression of thousands of genes by changes in a few functional processes [141] or in transcription factor activity [155]. The binary matrix encodes which genes participate in which functional processes or which genes are regulated by which transcription factors.

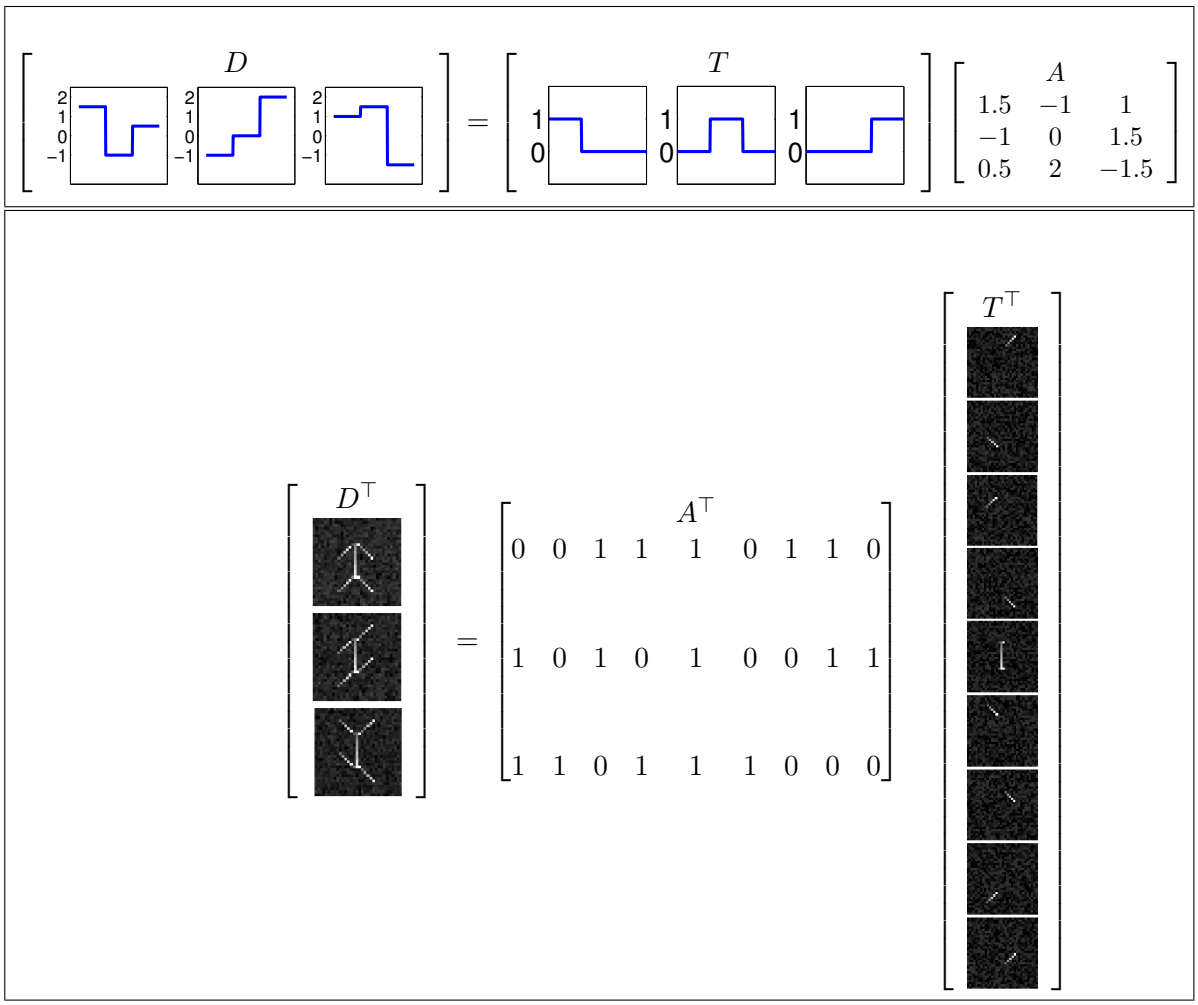


Figure 2.1: Illustration of the two matrix factorizations (2.7) and (2.10). Top: Each data point corresponds to a column of the matrix D, which can be represented as an affine combination of binary components. Bottom: Each data point corresponds to a row of $D^\top$, and the binary matrix represents presence/absence of the 'parts' contained in T. For example, the first row contains parts 3,4,5,6,7. All rows contain part 5. The bottom illustration has been borrowed from the swimmer data set in [48].

- In [4] a factorization of the form (2.9) is considered where the binary matrix is interpreted as a cluster membership matrix of a clustering where each data point may be assigned to multiple clusters. In this context, we point out that k-means clustering can be rephrased as the following matrix factorization (cf. e.g. [41]):

$$D \approx TA, \quad T \in \mathbb{R}^{m\times k}, \quad A \in \{0,1\}^{k\times n}, \quad A^\top \mathbf{1}_k = \mathbf{1}_n. \tag{2.11}$$

  It is important to note that in (2.11), the sum-to-one constraints are on the binary instead of the non-binary factor. Thus, our framework does not cover the k-means problem.

Several other matrix factorizations involving binary matrices have been proposed in the literature. In [83, 136] matrix factorization for binary input data, but non-binary

factors T and A is discussed. In [107] a factorization TWA with both T and A binary and real-valued W is proposed, which is more restrictive than the model of the present paper. The model in [107] in turn encompasses binary matrix factorization as proposed in [179], where all of D, T and A are constrained to be binary. It is important to note that this line of research is fundamentally different from Boolean matrix factorization [118], which is sometimes also referred to as binary matrix factorization.

2.4.2 Contributions

- As discussed above, structured matrix factorization problems are computationally challenging because of non-convexity. Even after relaxing the $\{0,1\}$-constraints into box constraints, it is not clear how to obtain a globally optimal solution. The combinatorial constraints seem to yield a further obstacle. Despite the obvious hardness of the problem, we present as our main contribution an algorithm that provably provides an exact factorization $D = TA$ with T and A as in (2.7) whenever such factorization exists. Our algorithm has a runtime complexity of $O(mr2^r + mnr + r^2n)$ flops, which is exponential in r but only linear in m and n. In particular, the problem remains tractable even for large values of m as long as r remains small.
- We show empirically that integer linear programming can be employed to achieve a dramatic speed-up of our algorithm for larger values of r. This allows us to solve problems with r as large as 80 ($2^{80} \approx 10^{24}$).
- We establish uniqueness of the exact factorization under separability (Definition 2.3), or alternatively with high probability for T drawn uniformly at random. As a corollary, we obtain that at least for these two models, the suggested algorithm continues to be fully applicable if additional constraints e.g. non-negativity, are imposed on the right factor A, cf. (2.8).
- Both the use of integer linear programming and the question of uniqueness are linked to a rich theory revolving around a central result in combinatorics, the Littlewood-Offord lemma.
- We extend the proposed algorithm to the approximate case $D \approx TA$ and empirically show superior performance relative to heuristic approaches to the problem.

2.4.3 Exact case

Given $D \in \mathbb{R}^{m\times n}$, we consider the following problem.

$$\text{find } T \in \{0,1\}^{m\times r} \text{ and } A \in \mathbb{R}^{r\times n},\ A^\top \mathbf{1}_r = \mathbf{1}_n \text{ such that } D = TA. \tag{2.12}$$

The columns $\{T_{:,k}\}_{k=1}^r$ of T can be identified with vertices of the hypercube $[0,1]^m$. The constraint $A^\top \mathbf{1}_r = \mathbf{1}_n$ is imposed for reasons of presentation, in order to avoid that the origin is treated differently from the other vertices of $[0,1]^m$, because otherwise the zero vector could be dropped from T, leaving the factorization unchanged. The additional constraint is not essential to our approach, see §2.4.4 below. We further

assume w.l.o.g. that r is minimal, i.e. there is no factorization of the form (2.12) with $r' < r$, and in turn that the columns of T are affinely independent, i.e. $\forall \lambda \in \mathbb{R}^r$, $\lambda^\top \mathbf{1}_r = 0$, $T\lambda = 0$ implies that $\lambda = 0$. Moreover, it is assumed that $\operatorname{rank}(A) = r$. This ensures the existence of a submatrix $A_{:,\mathcal{C}}$ of r linearly independent columns and of a corresponding submatrix of $D_{:,\mathcal{C}}$ of affinely independent columns, when combined with the affine independence of the columns of T:

$$\forall \lambda \in \mathbb{R}^r,\ \lambda^\top \mathbf{1}_r = 0 :\ D_{:,\mathcal{C}}\lambda = 0 \iff T(A_{:,\mathcal{C}}\lambda) = 0 \implies A_{:,\mathcal{C}}\lambda = 0 \implies \lambda = 0, \quad (2.13)$$

using at the second step that $\mathbf{1}_r^\top A_{:,\mathcal{C}}\lambda = \mathbf{1}_r^\top \lambda = 0$ and the affine independence of the $\{T_{:,k}\}_{k=1}^r$. Note that the assumption $\operatorname{rank}(A) = r$ is natural; otherwise, the data would reside in an affine subspace of lower dimension so that D would not contain enough information to reconstruct T.

2.4.4 Approach

Property (2.13) already provides the entry point of our approach. From $D = TA$, it is obvious that $\operatorname{aff}(T) \supseteq \operatorname{aff}(D)$. Since D contains the same number of affinely independent columns as T, it must also hold that $\operatorname{aff}(D) \supseteq \operatorname{aff}(T)$, in particular $\operatorname{aff}(D) \supseteq \{T_{:,k}\}_{k=1}^r$. Consequently, (2.12) can in principle be solved by enumerating all vertices of $[0,1]^m$ contained in $\operatorname{aff}(D)$ and selecting a maximal affinely independent subset thereof (see Figure 2.2 for an illustration). This procedure, however, is exponential in the dimension m, with 2^m vertices to be checked for containment in $\operatorname{aff}(D)$ by solving a linear system. Remarkably, the following observation along with its proof, which prompts Algorithm 2.2 below, shows that the number of elements to be checked can be reduced to 2^{r-1} irrespective of m.

Algorithm 2.2 FindVertices exact

1. Fix $p \in \operatorname{aff}(D)$ and compute $P = [D_{:,1} - p, \ldots, D_{:,n} - p]$.
2. Determine $r-1$ linearly independent columns $\mathcal{C}$ of P, obtaining $P_{:,\mathcal{C}}$ and subsequently $r-1$ linearly independent rows $\mathcal{R}$, obtaining $P_{\mathcal{R},\mathcal{C}} \in \mathbb{R}^{(r-1)\times(r-1)}$.
3. Form $Z = P_{:,\mathcal{C}}(P_{\mathcal{R},\mathcal{C}})^{-1} \in \mathbb{R}^{m\times(r-1)}$ and $\widehat{T} = Z(B^{(r-1)} - p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top) + p\mathbf{1}_{2^{r-1}}^\top \in \mathbb{R}^{m\times 2^{r-1}}$, where the columns of $B^{(r-1)}$ correspond to the elements of $\{0,1\}^{r-1}$.
4. Set $\mathcal{T} = \emptyset$. For $u = 1, \ldots, 2^{r-1}$, if $\widehat{T}_{:,u} \in \{0,1\}^m$ set $\mathcal{T} = \mathcal{T} \cup \{\widehat{T}_{:,u}\}$.
5. Return $\mathcal{T} = \{0,1\}^m \cap \operatorname{aff}(D)$.

Proposition 2.4. *The affine subspace aff(D) contains no more than 2^{r-1} vertices of $[0,1]^m$. Moreover, Algorithm 2.2 provides all vertices contained in aff(D).*

Proof. Consider the first part of the statement. Let $b \in \{0,1\}^m$ and $p \in \operatorname{aff}(D)$ arbitrary. We have $b \in \operatorname{aff}(D)$ iff there exists $\theta \in \mathbb{R}^n$ s.t.

$$D\theta = b,\ \theta^\top \mathbf{1}_n = 1 \iff \underbrace{[D_{:,1} - p, \ldots, D_{:,n} - p]}_{=P}\theta + p = b \iff P\theta = b - p. \quad (2.14)$$

Note that $\operatorname{rank}(P) = r - 1$. Hence, if there exists θ s.t. $P\theta = b - p$, such θ can be obtained from the unique $\lambda \in \mathbb{R}^{r-1}$ solving $P_{\mathcal{R},\mathcal{C}}\lambda = b_{\mathcal{R}} - p_{\mathcal{R}}$, where $\mathcal{R} \subset \{1, \ldots, m\}$ and $\mathcal{C} \subset \{1, \ldots, n\}$ are subsets of rows respectively columns of P s.t. $\operatorname{rank}(P_{\mathcal{R},\mathcal{C}}) = r - 1$. Finally note that $b_{\mathcal{R}} \in \{0,1\}^{r-1}$ so that there are no more than 2^{r-1} distinct right hand sides $b_{\mathcal{R}} - p_{\mathcal{R}}$.
Turning to the second part of the statement, observe that for each $b \in \{0,1\}^m$, there exists a unique λ s.t. $P_{\mathcal{R},\mathcal{C}}\lambda = b_{\mathcal{R}} - p_{\mathcal{R}} \Leftrightarrow \lambda = (P_{\mathcal{R},\mathcal{C}})^{-1}(b_{\mathcal{R}} - p_{\mathcal{R}})$. Repeating the argument preceding (2.14), if $b \in \{0,1\}^m \cap \operatorname{aff}(D)$, it must hold that

$$b = P_{:,\mathcal{C}}\lambda + p \iff b = \underbrace{P_{:,\mathcal{C}}(P_{\mathcal{R},\mathcal{C}})^{-1}}_{=Z}(b_{\mathcal{R}} - p_{\mathcal{R}}) + p \iff b = Z(b_{\mathcal{R}} - p_{\mathcal{R}}) + p. \quad (2.15)$$

Algorithm 2.2 generates all possible right hand sides $\widehat{T} = Z(B^{(r-1)} - p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top) + p\mathbf{1}_{2^{r-1}}^\top$, where $B^{(r-1)}$ contains all elements of $\{0,1\}^{r-1}$ as its columns. Consequently, if $b \in \{0,1\}^m \cap \operatorname{aff}(D)$, it must appear as a column of $\widehat{T}$. Conversely, if the leftmost equality in (2.15) does not hold, $b \notin \operatorname{aff}(D)$ and the column of $\widehat{T}$ corresponding to $b_{\mathcal{R}}$ cannot be a binary vector. □

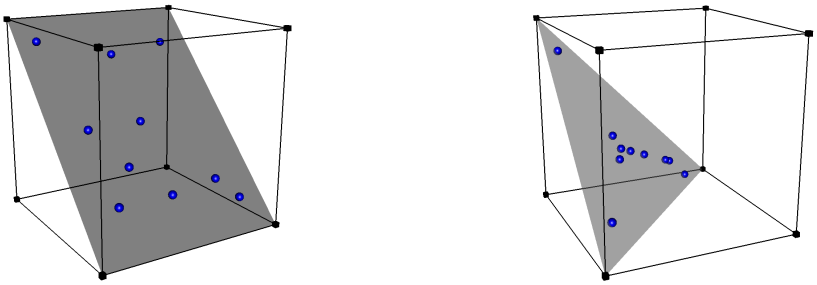


Figure 2.2: Illustration of the geometry underlying our approach in dimension $m = 3$. Dots represent data points and the shaded areas their affine hulls $\operatorname{aff}(D) \cap [0,1]^m$. Left: $\operatorname{aff}(D)$ intersects with $r + 1$ vertices of $[0,1]^m$. Right: $\operatorname{aff}(D)$ intersects with precisely r vertices.

Algorithm 2.3 BinaryFactorization exact

1. Obtain $\mathcal{T}$ as output from FindVertices Exact(D)
2. Select r affinely independent elements of $\mathcal{T}$ to be used as columns of T.
3. Obtain A as solution of the linear system $[\mathbf{1}_r^\top; T]A = [\mathbf{1}_n^\top; D]$.
4. Return (T, A) solving problem (2.12).

Comments. In step 2 of Algorithm 2.2, determining the rank of P and an associated set of linearly independent columns/rows can be done by means of a rank-revealing QR factorization [69, 72]. The crucial step is the third one, which is a compact description of first solving the linear systems $P_{\mathcal{R},\mathcal{C}}\lambda = b - p_{\mathcal{R}}$ for all $b \in \{0,1\}^{r-1}$ and back-substituting

the result to compute candidate vertices $P_{:,\mathcal{C}}\lambda + p$ stacked into the columns of $\widehat{T}$; the addition/subtraction of p is merely because we have to deal with an affine instead of a linear subspace, in which p serves as origin. In step 4, the pool of 2^{r-1} 'candidates' is filtered, yielding $\mathcal{T} = \mathrm{aff}(D) \cap \{0,1\}^m$.
Determining $\mathcal{T}$ is the hardest part in solving the matrix factorization problem (2.12). Given $\mathcal{T}$, the solution can be obtained after few inexpensive standard operations. Note that step 2 in Algorithm 2.3 is not necessary if one does not aim at finding a minimal factorization, i.e. if it suffices to have $D = TA$ with $T \in \{0,1\}^{m\times r'}$ but r' possibly being larger than r.

Computational complexity. The dominating cost in Algorithm 2.2 is computation of the candidate matrix $\widehat{T}$ and checking whether its columns are vertices of $[0,1]^m$. Note that

$$\widehat{T}_{\mathcal{R},:} = Z_{\mathcal{R},:}(B^{(r-1)} - p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top) + p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top = I_{r-1}(B^{(r-1)} - p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top) + p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top = B^{(r-1)}, \tag{2.16}$$

i.e. the $r-1$ rows of $\widehat{T}$ corresponding to $\mathcal{R}$ do not need to be taken into account. Forming the matrix $\widehat{T}$ would hence require $O((m-r+1)(r-1)2^{r-1})$ and the subsequent check for vertices in the fourth step $O((m-r+1)2^{r-1})$ operations. All other operations are of lower order provided e.g. $(m-r+1)2^{r-1} > n$. The second most expensive operation is forming the matrix $P_{\mathcal{R},\mathcal{C}}$ in step 2 with the help of a QR decomposition requiring $O(mn(r-1))$ operations in typical cases [72]. Computing the matrix factorization (2.12) after the vertices have been identified (steps 2 to 4 in Algorithm 2.3) has complexity $O(mnr + r^3 + r^2 n)$. Here, the dominating part is the solution of a linear system in r variables and n right hand sides. Altogether, our approach for solving (2.12) has exponential complexity in r, but only linear complexity in m and n. Later on, we will argue that under additional assumptions on T, the $O((m-r+1)2^{r-1})$ terms can be reduced to $O((r-1)2^{r-1})$.

Variants.

(1) We point out that the approach discussed above remains applicable after simple changes if $\{0,1\}$ is replaced e.g. by $\{-1,1\}$ or $\{0.1,0.9\}$. This amounts to scaling and translation of $[0,1]^m$, which does not conceptually affect our approach. One could consider as well sets of size q, e.g. for $q=3$, $\{0,0.5,1\}$, in which case one would check q^{r-1} candidates in Algorithm 2.2.

(2) We here provide variants of Algorithms 2.2 and 2.3 to solve the corresponding problem without the constraint $A^\top \mathbf{1}_r = \mathbf{1}_n$, that is

$$\text{find } T \in \{0,1\}^{m\times r} \text{ and } A \in \mathbb{R}^{r\times n} \text{ such that } D = TA. \tag{2.17}$$

The above Algorithm 2.4 is the analog of Algorithm 2.2. Algorithm 2.4 yields $\mathrm{span}(D)\cap \{0,1\}^m$, which can be proved along the lines of the proof of Proposition 2.4 under the stronger assumption that T has r *linearly* independent in place of only r *affinely* independent columns, which together with the assumption $\mathrm{rank}(A) = r$ implies that

Algorithm 2.4 FindVertices exact_linear

1. Determine r linearly independent columns $\mathcal{C}$ of D, obtaining $D_{:,\mathcal{C}}$ and subsequently r linearly independent rows $\mathcal{R}$, obtaining $D_{\mathcal{R},\mathcal{C}} \in \mathbb{R}^{r\times r}$.
2. Form $Z = D_{:,\mathcal{C}}(D_{\mathcal{R},\mathcal{C}})^{-1} \in \mathbb{R}^{m\times r}$ and $\widehat{T} = ZB^{(r)} \in \mathbb{R}^{m\times 2^r}$, where the columns of $B^{(r)}$ correspond to the elements of $\{0,1\}^r$
3. Set $\mathcal{T} = \emptyset$. For $u = 1,\ldots,2^r$, if $\widehat{T}_{:,u} \in \{0,1\}^m$ set $\mathcal{T} = \mathcal{T} \cup \{\widehat{T}_{:,u}\}$.
4. Return $\mathcal{T} = \{0,1\}^m \cap \operatorname{span}(D)$.

also $\operatorname{rank}(D) = r$. Algorithm 2.4 results from Algorithm 2.2 by setting $p = 0$ and replacing $r-1$ by r.
The following Algorithm 2.5 solves problem (2.17) given the output of Algorithm 2.4.

Algorithm 2.5 BinaryFactorization exact_linear

1. Obtain $\mathcal{T}$ as output from FindVertices Exact_linear(D)
2. Select r linearly independent elements of $\mathcal{T}$ to be used as columns of T.
3. Obtain A as solution of the linear system $TA = D$.
4. Return (T, A) solving problem (2.17).

(3) We here sketch how our approach can be applied to the following matrix factorization problem considered in [107].

$$\text{find } T \in \{0,1\}^{m\times r},\ A \in \{0,1\}^{n\times r} \text{ and } W \in \mathbb{R}^{r\times r} \text{ such that } D = TWA^\top, \tag{2.18}$$

Suppose that $\operatorname{rank}(D) = r$. Then the following Algorithm 2.6 solves problem (2.18). Since Algorithm 2.6 can be reduced to a twofold application of Algorithm (2.17), the proof is omitted.

Algorithm 2.6 ThreeWayBinaryFactorization

1. Obtain $\mathcal{T}$ as output from FindVertices Exact_linear(D)
2. Obtain $\mathcal{A}$ as output from FindVertices Exact_linear($D^\top$)
3. Select r linearly independent elements of $\mathcal{T}$ and $\mathcal{A}$ to be used as columns of T respectively A.
4. Obtain $W = (T^\top T)^{-1}T^\top DA(A^\top A)^{-1}$.

2.4.5 Uniqueness

In this subsection, we study uniqueness of the matrix factorization problem (2.12) (modulo permutation of columns/rows). First note that in view of the affine independence of the columns of T, the factorization is unique iff T is, which holds iff

$$\operatorname{aff}(D) \cap \{0,1\}^m = \operatorname{aff}(T) \cap \{0,1\}^m = \{T_{:,1},\ldots,T_{:,r}\}, \tag{2.19}$$

i.e. if the affine subspace generated by $\{T_{:,1}, \dots, T_{:,r}\}$ contains no other vertices of $[0,1]^m$ than the r given ones (cf. Figure 2.2). Uniqueness is of great importance in applications, where one aims at an interpretation in which the columns of T play the role of underlying data-generating elements. Such an interpretation is not valid if (2.19) fails to hold, since it is then possible to replace one of the columns of a specific choice of T by another vertex contained in the same affine subspace.

Solution of a non-negative variant of our factorization. In the sequel, we argue that property (2.19) plays an important role from a computational point of view when solving extensions of problem (2.12) in which further constraints are imposed on A. One particularly important extension is the following.

$$\text{find } T \in \{0,1\}^{m\times r} \text{ and } A \in \mathbb{R}_+^{r\times n},\ A^\top \mathbf{1}_r = \mathbf{1}_n \text{ such that } D = TA. \tag{2.20}$$

Problem (2.20) is a special instance of NMF. The additional non-negativity constraints are of particular interest here, because they arise in the real world application which has motivated this work; see §2.4.8 below. It is natural to ask whether Algorithm 2.3 can be adapted to solve problem (2.20). A change is obviously required for the second step when selecting r vertices from $\mathcal{T}$, since in (2.20) the columns D now have to be expressed as convex instead of only affine combinations of columns of T: picking an affinely independent collection from $\mathcal{T}$ does not take into account the non-negativity constraint imposed on A. If, however, (2.19) holds, we have $|\mathcal{T}| = r$ and Algorithm 2.3 must return a solution of (2.20) provided that there exists one.

Corollary 2.5. *If problem* (2.12) *has a unique solution, i.e. if condition* (2.19) *holds and if there exists a solution of* (2.20)*, then it is returned by Algorithm 2.3.*

Corollary 2.5 follows immediately from Proposition 2.4. Note that analogous results hold for *arbitrary* (not necessarily convex) constraints imposed on A. In this sense, the statement can be seen as trivial. Nevertheless, to appreciate that result, consider the converse case $|\mathcal{T}| > r$. Since the aim is a minimal factorization, one has to find a subset of $\mathcal{T}$ of cardinality r such that (2.20) can be solved. In principle, this can be achieved by considering all $\binom{|\mathcal{T}|}{r}$ subsets of $\mathcal{T}$, but this is in general not computationally feasible: the upper bound of Proposition 2.4 indicates that $|\mathcal{T}| = 2^{r-1}$ in the worst case. For the example below, $\mathcal{T}$ consists of all 2^{r-1} vertices contained in an $r-1$-dimensional face of $[0,1]^m$:

$$T = \begin{pmatrix} 0_{(m-r)\times r} \\ I_{r-1}\ 0_{r-1} \\ 0_r^\top \end{pmatrix} \quad \text{with } \mathcal{T} = \left\{ T\lambda :\ \lambda_1 \in \{0,1\}, \dots, \lambda_{r-1} \in \{0,1\},\ \lambda_r = 1 - \sum_{k=1}^{r-1} \lambda_k \right\}. \tag{2.21}$$

Remark. In Appendix B, we show that in general even the NMF problem (2.20) may not have a unique solution (note that failure of (2.19) only implies non-uniqueness of problem (2.12)). While it is well-known that NMF need not to have a unique solution [48], it is rather remarkable that even with binary constraints on one factor, which is a strong additional restriction, uniqueness may fail.

Uniqueness under separability. In view of the negative example (2.21), one might ask whether uniqueness according to (2.19) can at least be achieved under additional conditions on T. Below we prove uniqueness under separability (Definition 2.3).

Proposition 2.6. *If T is separable, condition (2.19) holds and thus problem (2.12) has a unique solution.*

Proof. We have $\mathrm{aff}(T) \ni b \in \{0,1\}^m$ iff there exists $\lambda \in \mathbb{R}^r$, $\lambda^\top \mathbf{1}_r = 1$ such that $T\lambda = b$. Since T is separable, there exists a permutation matrix Π such that $\Pi T = [I_r; M]$ with $M \in \{0,1\}^{(m-r)\times r}$. As a result,

$$T\lambda = b \iff \Pi T\lambda = \Pi b \iff [I_r; M]\lambda = \Pi b.$$

Since $\Pi b \in \{0,1\}^m$, for the top r block of the linear system to be fulfilled, it is necessary that $\lambda \in \{0,1\}^r$. The condition $\lambda^\top \mathbf{1}_r = 1$ then implies that λ must be one of the r canonical basis vectors of $\mathbb{R}^r$. We conclude that $\mathrm{aff}(T) \cap \{0,1\}^m = \{T_{:,1}, \ldots, T_{:,r}\}$. □

Uniqueness under generic random sampling. Both the negative example (2.21) as well as the positive result of Proposition 2.6 are associated with special matrices T. This raises the question whether uniqueness holds respectively fails for broader classes of binary matrices. In order to gain insight into this question, we consider random T with i.i.d. entries from a Bernoulli distribution with parameter $\frac{1}{2}$ and study the probability of the event $\{\mathrm{aff}(T) \cap \{0,1\}^m = \{T_{:,1}, \ldots, T_{:,r}\}\}$. This question has essentially been studied in combinatorics [117], with further improvements in [84]. The results therein rely crucially on Littlewood-Offord theory, a topic we will touch upon in the subsequent paragraph.

Theorem 2.7. *Let T be a random $m \times r$-matrix whose entries are drawn i.i.d. from $\{0,1\}$ with probability $\frac{1}{2}$. Then, there is a constant C so that if $r \le m - C$,*

$$\mathbf{P}\left(\mathit{aff}(T) \cap \{0,1\}^m = \{T_{:,1}, \ldots, T_{:,r}\}\right) \ge 1 - (1 + o(1))\, 4\binom{r}{3}\left(\frac{3}{4}\right)^m - \left(\frac{3}{4} + o(1)\right)^m$$

as $m \to \infty$.

Our proof of Theorem 2.7 relies on on two seminal results on random ± 1-matrices.

Theorem 2.8. *[84] Let M be a random $m \times r$-matrix whose entries are drawn i.i.d. from $\{-1,1\}$ each with probability $\frac{1}{2}$. There is a constant C so that if $r \le m - C$,*

$$\mathbf{P}\left(\mathit{span}(M) \cap \{-1,1\}^m = \{\pm M_{:,1}, \ldots, \pm M_{:,r}\}\right) \ge 1 - (1 + o(1))\, 4\binom{r}{3}\left(\frac{3}{4}\right)^m \quad (2.22)$$

as $m \to \infty$.

Theorem 2.9. *[148] Let M be a random $m \times r$-matrix, $r \le m$, whose entries are drawn i.i.d. from $\{-1,1\}$ each with probability $\frac{1}{2}$. Then*

$$\mathbf{P}\left(M \text{ has linearly independent columns}\right) \ge 1 - \left(\frac{3}{4} + o(1)\right)^m \quad \text{as } m \to \infty. \quad (2.23)$$

Proof. (Theorem 2.7) Note that $T = \frac{1}{2}(M + \mathbf{1}_{m\times r})$, where M is a random ± 1-matrix as in Theorem 2.8. Let $\lambda \in \mathbb{R}^r$, $\lambda^\top \mathbf{1}_r = 1$ and $b \in \{0,1\}^m$. Then

$$T\lambda = b \iff \frac{1}{2}(M\lambda + \mathbf{1}_m) = b \iff M\lambda = 2b - \mathbf{1}_m \in \{-1,1\}^m. \tag{2.24}$$

Now note that with the probability given in (2.22),

$$\begin{aligned} &\operatorname{span}(M) \cap \{-1,1\}^m = \{\pm M_{:,1}, \ldots, \pm M_{:,r}\} \\ \Longrightarrow\ &\operatorname{aff}(M) \cap \{-1,1\}^m \subseteq \{\pm M_{:,1}, \ldots, \pm M_{:,r}\} \end{aligned}$$

On the other hand, with the probability given in (2.23), the columns of M are linearly independent. If this is the case,

$$\begin{aligned} &\operatorname{aff}(M) \cap \{-1,1\}^m \subseteq \{\pm M_{:,1}, \ldots, \pm M_{:,r}\} \\ \Longrightarrow\ &\operatorname{aff}(M) \cap \{-1,1\}^m = \{M_{:,1}, \ldots, M_{:,r}\}. \end{aligned} \tag{2.25}$$

To verify this, first note the obvious inclusion $\operatorname{aff}(M) \cap \{-1,1\}^m \supseteq \{M_{:,1}, \ldots, M_{:,r}\}$. Moreover, suppose by contradiction that there exists $j \in \{1, \ldots, r\}$ and $\theta \in \mathbb{R}^r$, $\theta^\top \mathbf{1}_r = 1$ such that $M\theta = -M_{:,j}$. Writing e_j for the j-th canonical basis vector, this would imply $M(\theta + e_j) = 0$ and in turn by linear independence $\theta = -e_j$, which contradicts $\theta^\top \mathbf{1}_r = 1$.
Under the event (2.25), $M\lambda = 2b - \mathbf{1}_m$ is fulfilled iff λ is equal to one of the canonical basis vectors and $2b - \mathbf{1}_m$ equals the corresponding column of M. We conclude the assertion in view of (2.24). □

Theorem 2.7 suggests a positive answer to the question of uniqueness posed above. Asymptotically as $m \to \infty$ and for r small compared to m (in fact, following [84] one may conjecture that Theorem 2.7 holds with $C = 1$), the probability that the affine hull of r vertices of $[0,1]^m$ selected uniformly at random contains some other vertex is exponentially small in the dimension m. It is natural to ask whether a result similar to Theorem 2.7 holds if the entries of T are drawn from a Bernoulli distribution with parameter p in $(0,1)$ sufficiently far away from the boundary points. Second, it is of interest to know whether the above statement is already valid for finite, though reasonably large values of m. We have therefore conducted an experiment whose outcome suggests that the answers to both questions are positive. For this experiment, we consider the grid $\{0.01, 0.02, \ldots, 0.99\}$ for p and generate random binary matrices $T \in \mathbb{R}^{m\times r}$ with $m = 500$ and $r \in \{8, 16, 24\}$ whose entries are i.i.d. Bernoulli with parameter p. For each value of p and r, 100 trials are considered, and for each of these trials, we compute the number of vertices of $[0,1]^m$ contained in $\operatorname{aff}(T)$. In Figure 2.3, we report the maximum number of vertices over these trials. One observes that except for a small set of values of p very close to 0 or 1, exactly r vertices are returned in all trials. On the other hand, for extreme values of p the number of vertices can be as large as 2^{20} in the worst case.
As a byproduct, these results indicate that also the NMF variant of our matrix factorization problem (2.20) can in most cases be reduced to identifying a set of r vertices of $[0,1]^m$ (cf. Corollary 2.5).

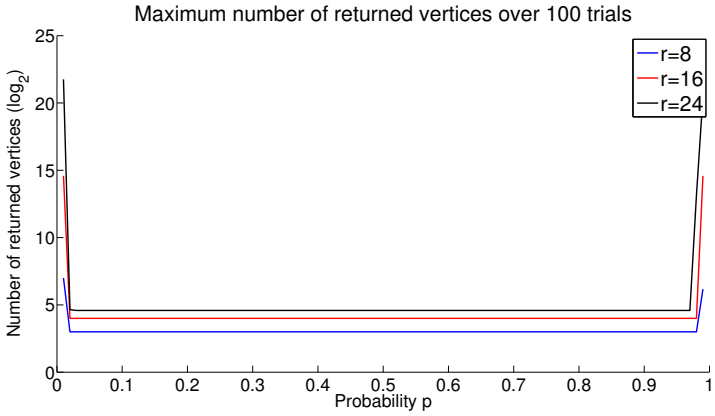


Figure 2.3: Number of vertices contained in $\mathrm{aff}(T)$ over 100 trials for T drawn entry-wise from a Bernoulli distribution with parameter p.

2.4.6 Speeding up the basic algorithm

In Algorithm 2.2, an $m \times 2^{r-1}$ matrix $\widehat{T}$ of potential vertices is formed (step 3). We have discussed the case (2.21) where all candidates must indeed be vertices, in which case it seems impossible to reduce the computational cost of $O((m-r)r2^{r-1})$, which becomes significant once m is in the thousands and $r \geq 25$. On the positive side, Theorem 2.7 indicates that for many instances of T, only r out of 2^{r-1} candidates are in fact vertices. In that case, noting that columns of $\widehat{T}$ cannot be vertices if a single coordinate is not in $\{0,1\}$ (and that the vast majority of columns of $\widehat{T}$ must have such coordinate), it is computationally more favourable to incrementally compute subsets of rows of $\widehat{T}$ and then to discard already those columns with coordinates not in $\{0,1\}$. We have found empirically that this scheme rapidly reduces the candidate set − already checking a single row of $\widehat{T}$ eliminates a substantial portion (see Figure 2.4).

Littlewood-Offord theory. Theoretical underpinning for the last observation can be obtained from a result in combinatorics, the Littlewood-Offord (L-O) lemma. Various extensions of that result have been developed until recently, see the survey [116]. We here cite the L-O lemma in its basic form.

Theorem 2.10. *[57] Let $a_1, \ldots, a_\ell \in \mathbb{R} \setminus \{0\}$ and $y \in \mathbb{R}$.*

(i) $\left|\{b \in \{0,1\}^\ell : \sum_{i=1}^{\ell} a_i b_i = y\}\right| \leq \binom{\ell}{\lfloor \ell/2 \rfloor}$.

(ii) If $|a_i| \geq 1$, $i = 1, \ldots, \ell$, $\left|\{b \in \{0,1\}^\ell : \sum_{i=1}^{\ell} a_i b_i \in (y, y+1)\}\right| \leq \binom{\ell}{\lfloor \ell/2 \rfloor}$.

The two parts of Theorem 2.10 are referred to as discrete respectively continuous L-O lemma. The discrete L-O lemma provides an upper bound on the number of $\{0,1\}$-vectors whose weighted sum with given weights $\{a_i\}_{i=1}^{\ell}$ is equal to some given number y, whereas the stronger continuous version, under a more stringent condition on the weights, upper bounds the number of $\{0,1\}$-vectors whose weighted sum is contained in

some interval $(y, y+1)$. In order to see the relation of Theorem 2.10 to Algorithm 2.2, let us re-inspect the third step of that algorithm. To obtain a reduction of candidates by checking a single row of $\widehat{T} = Z(B^{(r-1)} - p_{\mathcal{R}}\mathbf{1}_{2^{r-1}}^\top) + p\mathbf{1}_{2^{r-1}}^\top$, pick $i \notin \mathcal{R}$ (recall that coordinates in $\mathcal{R}$ do not need to be checked, cf. (2.16)) and $u \in \{1, \ldots, 2^{r-1}\}$ arbitrary. The u-th candidate can be a vertex only if $\widehat{T}_{i,u} \in \{0, 1\}$. The condition $\widehat{T}_{i,u} = 0$ can be written as

$$\underbrace{Z_{i,:}}_{\{a_k\}_{k=1}^r} \underbrace{B_{:,u}^{(r-1)}}_{=b} = \underbrace{Z_{i,:}p_{\mathcal{R}} - p_i}_{=y}. \tag{2.26}$$

A similar reasoning applies when setting $\widehat{T}_{i,u} = 1$. Provided that none of the entries of $Z_{i,:} = 0$, the discrete L-O lemma implies that there are at most $2\binom{r-1}{\lfloor (r-1)/2 \rfloor}$ out of 2^{r-1} candidates for which the i-th coordinate is in $\{0, 1\}$. This yields a reduction of the candidate set by $2\binom{r-1}{\lfloor (r-1)/2 \rfloor}/2^{r-1} = O\left(\frac{1}{\sqrt{r-1}}\right)$. Admittedly, this reduction may appear insignificant given the total number of candidates to be checked. The reduction achieved empirically (cf. Figure 2.4) is typically larger. Stronger reductions have been proven under additional assumptions on the weights $\{a_i\}_{i=1}^{\ell}$: e.g. for distinct weights, one obtains a reduction of $O((r-1)^{-3/2})$ [116]. Furthermore, when picking successively d rows of $\widehat{T}$ and if one assumes that each row yields a reduction according to the discrete L-O lemma, one would obtain the reduction $(r-1)^{-d/2}$ so that $d = r - 1$ would suffice to identify all vertices provided $r \geq 5$. Evidence for the rate $(r-1)^{-d/2}$ can be found in [149]. This indicates a reduction in complexity of Algorithm 2.2 from $O((m-r)r2^{r-1})$ to $O(r^2 2^{r-1})$.

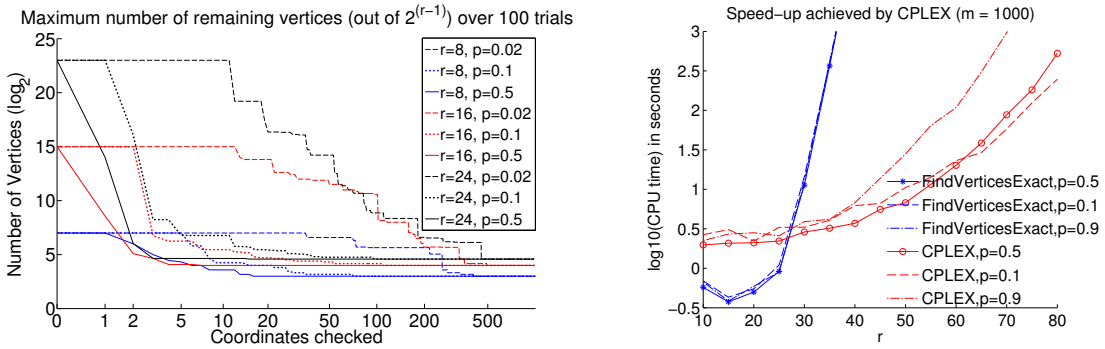


Figure 2.4: Left: Speeding up the algorithm by checking single coordinates, remaining number of coordinates vs.# coordinates checked ($m = 1000$). Right: Speed up by CPLEX compared to Algorithm 2.2. For both plots, T is drawn entry-wise from a Bernoulli distribution with parameter p.

Achieving further speed-up with integer linear programming. The continuous L-O lemma (part (ii) of Theorem 2.10) combined with the derivation leading to (2.26) suggests an approach that allows us to tackle even the case $r = 80$ ($2^{80} \approx 10^{24}$). In view of the continuous L-O lemma, a reduction in the number of candidates may still be achievable if the requirement is weakened to $\widehat{T}_{i,u} \in [0, 1]$. According to (2.26) the candidates satisfying the relaxed constraint for the i-th coordinate can be obtained

from the feasibility problem

$$\text{find } b \in \{0,1\}^{r-1} \text{ subject to } 0 \leq Z_{i,:}(b - p_{\mathcal{R}}) + p_i \leq 1, \tag{2.27}$$

which is an integer linear program that can be solved e.g. by CPLEX . The L-O theory suggests that the branch-and-cut strategy [170] employed therein is likely to be successful. With the help of CPLEX, it is affordable to solve problem (2.27) with all $m - r + 1$ constraints (one for each of the rows of $\widehat{T}$ to be checked) imposed simultaneously. We always recovered directly the underlying vertices in our experiments and only these, without the need to prune the solution pool (which could be achieved by Algorithm 2.2, replacing the 2^{r-1} candidates by a potentially much smaller solution pool).

2.4.7 Approximate case

In the sequel, we discuss an extension of our approach to handle the approximate case $D \approx TA$ with T and A as in (2.12). In particular, we have in mind the case of additive noise i.e. $D = TA + E$ with $\|E\|_F$ small. While the basic concept of Algorithm 2.2 can be adopted, changes are necessary because D may have full rank $\min\{m, n\}$ and second $\text{aff}(D) \cap \{0,1\}^m = \emptyset$, i.e. the distances of $\text{aff}(D)$ and the $\{T_{:,k}\}_{k=1}^r$ may be strictly positive (but are at least assumed to be small). As distinguished from the

Algorithm 2.7 FINDVERTICES APPROXIMATE

1. Let $p = D\mathbf{1}_n/n$ and compute $P = [D_{:,1} - p, \ldots, D_{:,n} - p]$.
2. Compute $U^{(r-1)} \in \mathbb{R}^{m \times (r-1)}$, the left singular vectors corresponding to the $r-1$ largest singular values of P [2]. Select $r-1$ linearly independent rows $\mathcal{R}$ of $U^{(r-1)}$, obtaining $U^{(r-1)}_{\mathcal{R},:} \in \mathbb{R}^{(r-1)\times(r-1)}$.
3. Form $Z = U^{(r-1)}(U^{(r-1)}_{\mathcal{R},:})^{-1}$ and $\widehat{T} = Z(B^{(r-1)} - p_{\mathcal{R}}\mathbf{1}^\top_{2^{r-1}}) + p\mathbf{1}^\top_{2^{r-1}}$.
4. Compute $\widehat{T}^{01} \in \mathbb{R}^{m \times 2^{r-1}}$: for $u = 1, \ldots, 2^{r-1}$, $i = 1, \ldots, m$, set $\widehat{T}^{01}_{i,u} = I(\widehat{T}_{i,u} > \frac{1}{2})$.
5. For $u = 1, \ldots, 2^{r-1}$, set $\delta_u = \|\widehat{T}_{:,u} - \widehat{T}^{01}_{:,u}\|_2$. Order increasingly s.t. $\delta_{u_1} \leq \ldots \leq \delta_{2^{r-1}}$.
6. Return $T = [\widehat{T}^{01}_{:,u_1} \ldots \widehat{T}^{01}_{:,u_r}]$

exact case, Algorithm 2.7 requires the number of components r to be specified in advance as it is typically the case in noisy matrix factorization problems. Moreover, the vector p subtracted from all columns of D in step 1 is chosen as the mean of the data points, which is in particular a reasonable choice if D is contaminated with additive noise distributed symmetrically around zero. The truncated SVD of step 2 achieves the desired dimension reduction and potentially reduces noise corresponding to small singular values that are discarded. The last change arises in step 5. While in the exact case, one identifies all columns of $\widehat{T}$ that are in $\{0,1\}^m$, one instead only identifies columns close to $\{0,1\}^m$. Given the output of Algorithm 2.7, we solve the approximate matrix factorization problem via (constrained) least squares, obtaining the right factor from $\min_{A \in \mathcal{C}_A} \|D - TA\|_F^2$, where the constraint set $\mathcal{C}_A = \{A \in \mathbb{R}^{r \times n} : A^\top \mathbf{1}_r = \mathbf{1}_n\}$

[2] cf. Theorem 2.2

or $\mathcal{C}_A = \{A \in \mathbb{R}_+^{r\times n} : A^\top \mathbf{1}_r = \mathbf{1}_n\}$ in the non-negative case. Again, the unit sum constraints can be dropped by modifying Algorithm 2.7 accordingly.

Refinements. Improved performance for higher noise levels can be achieved by running Algorithm 2.7 multiple times with different sets of rows selected in step 2, which yields candidate matrices $\{T^{(l)}\}_{l=1}^s$, and subsequently using the candidate yielding the best fit, i.e. one picks $T = \operatorname{argmin}_{\{T^{(l)}\}} \min_A \|D - T^{(l)}A\|_F^2$. Alternatively, we may form a candidate pool by merging the $\{T^{(l)}\}_{l=1}^s$ and then use a backward elimination scheme, in which successively candidates are dropped that yield the smallest improvement in fitting D until r candidates are left. Apart from that, T returned by Algorithm 2.7 can be used for initializing Algorithm 2.8 below, which falls under the block coordinate descent scheme of Algorithm 2.1.

Algorithm 2.8 Block optimization scheme for solving $\min_{T\in\{0,1\}^{m\times r},\, A\in\mathcal{C}_A} \|D - TA\|_F^2$

Initialize T^0
for $t = 0, 1, \ldots$ **do**

$$A^{t+1} \leftarrow \operatorname*{argmin}_{A\in\mathcal{C}_A} \|D - T^t A\|_F^2$$

$$T^{t+1} \leftarrow \operatorname*{argmin}_{T\in\{0,1\}^{m\times r}} \|D - TA^{t+1}\|_F^2 = \operatorname*{argmin}_{\{T_{i,:}\in\{0,1\}^r\}_{i=1}^m} \sum_{i=1}^m \|D_{i,:} - T_{i,:}A^{t+1}\|_2^2 \tag{2.28}$$

end for

An important observation is that optimization of T (2.28) is separable along the rows of T, so that for small r, it is feasible to perform exhaustive search over all 2^r possibilities (or to use CPLEX). However, Algorithm 2.8 is impractical as a stand-alone scheme, because without proper initialization, it may take many iterations to converge, with each single iteration being more expensive than Algorithm 2.7. When initialized with the output of the latter, however, we have observed convergence of the block optimization scheme only after few steps.

2.4.8 Experiments

In the first part, we demonstrate with the help of synthetic data that the approach of the preceding subsection performs well on noisy datasets. In the second part, we present an application to a real dataset.

Synthetic data.

Setups.
'T0.5': We generate $D = T^*A^* + \alpha E$, where the entries of T^* are drawn i.i.d. from $\{0, 1\}$ with probability 0.5, the columns of A are drawn i.i.d. uniformly from the probability simplex and the entries of E are i.i.d. standard Gaussian. We let $m = 1000$, $r \in \{10, 20\}$ and $n = 2r$ and let the noise level α vary along a grid starting from 0. Small sample sizes n as considered here yield more challenging problems and are motivated by the real world application of the subsequent paragraph.

'Tsparse+dense': The matrix T is now generated by drawing then entries of one half of the columns of T i.i.d. from a Bernoulli distribution with probability 0.1 ('sparse' part), and the second half from a Bernoulli distribution with parameter 0.9 ('dense' part). The rest is as for the first setup.

'T0.5,Adense' As for 'T0.5' apart from the following modification: after random generation of A as above, we compute its Euclidean projection on $\{A \in \mathbb{R}_+^{r\times n} : A^\top \mathbf{1}_r = \mathbf{1}_n,\ \max_{k,i} A_{k,i} \leq 2/r\}$, thereby constraining the columns of A to be roughly constant. With such A, all data points are situated near the barycentre $T\mathbf{1}_r/r$ of the simplex generated by the columns of T. Given that the goal is to recover vertices, this setup is hence potentially more difficult.

Methods compared.
FindVertices: Our approach as described in the previous subsection. After obtaining T as output from Algorithm 2.7, we solve the constrained least squares problem

$$\min_{\{A\in\mathbb{R}_+^{r\times n}:\, A^\top\mathbf{1}_r=\mathbf{1}_n\}} \|D - TA\|_F^2 = \min_{\{A_{:,j}\in\mathbb{R}_+^r,\, A_{:,j}^\top\mathbf{1}_r=1\}_{j=1}^n} \sum_{j=1}^n \|D_{:,j} - TA_{:,j}\|_F^2.$$

The optimization problem decouples along the columns of A, yielding a simplex-constrained least squares problem in r variables per columns, which we solve with spectral projected gradient [12] in conjunction with Michelot's algorithm [112] for computing the Euclidean projection on the simplex.

oracle. The oracle has access to the non-binary factor A^* and hence does not need solve a matrix factorization problem, but only a regression problem. Equipped with A^*, the oracle estimates T^* as

$$\operatorname*{argmin}_{T\in\{0,1\}^{m\times r}} \|D - TA^*\|_F^2 = \operatorname*{argmin}_{\{T_{i,:}\in\{0,1\}^r\}_{i=1}^m} \sum_{i=1}^m \|D_{i,:} - T_{i,:}A^*\|_2^2,$$

cf. (2.28) in Algorithm 2.8.

box. We relax the integer constraints into box constraints. This yields a structured matrix factorization problem of the form (2.4) with $\mathcal{C}_T = [0,1]^{m\times r}$ and $\mathcal{C}_A = \{A \in \mathbb{R}_+^{r\times n} : A^\top\mathbf{1}_r = \mathbf{1}_n\}$. Block coordinate descent (Algorithm 2.1) is used with five random initializations, and we take the result yielding the best fit out of these five trials. Subsequently, the entries of T are rounded to fulfill the $\{0,1\}$-constraints. This is followed by a re-fitting of A, i.e. one more update of A according to Algorithm 2.1 is performed. The block updates amount to least squares problems with simplex respectively box constraints, which are solved by spectral projected gradient.

quad pen. A slightly enhanced version of box in which a quad(ratic) pen(alty) is used to push the entries of T towards $\{0,1\}$. More specifically, we consider

$$\min_{\substack{T\in[0,1]^{m\times r}\\ A\in\mathbb{R}_+^{r\times n},\, A^\top\mathbf{1}_r=\mathbf{1}_n}} \|D - TA\|_F^2 + \lambda \sum_{i=1}^m \sum_{k=1}^r \omega(T_{ik}), \quad \text{where } \omega(t) = t(1-t). \tag{2.29}$$

Note that the quadratic function ω equals zero for $\{0, 1\}$ while it increases the more one moves away from the end points of the unit interval. For $\lambda > 0$ large enough, the set of minimizers of problem (2.29) coincides with the set of minimizers of the corresponding matrix factorization problem with binary constraints on T. Problem (2.29) can still be handled with a block coordinate descent scheme akin to Algorithm 2.1. However, because of the concavity of ω, already the update of T given A constitutes a non-convex problem. Since the associated objective is the sum of one convex and one concave term, we make use of the convex-concave procedure (CCCP, [172]), a popular algorithm for DC (difference of convex functions, [42]) programs; see Appendix B for details.

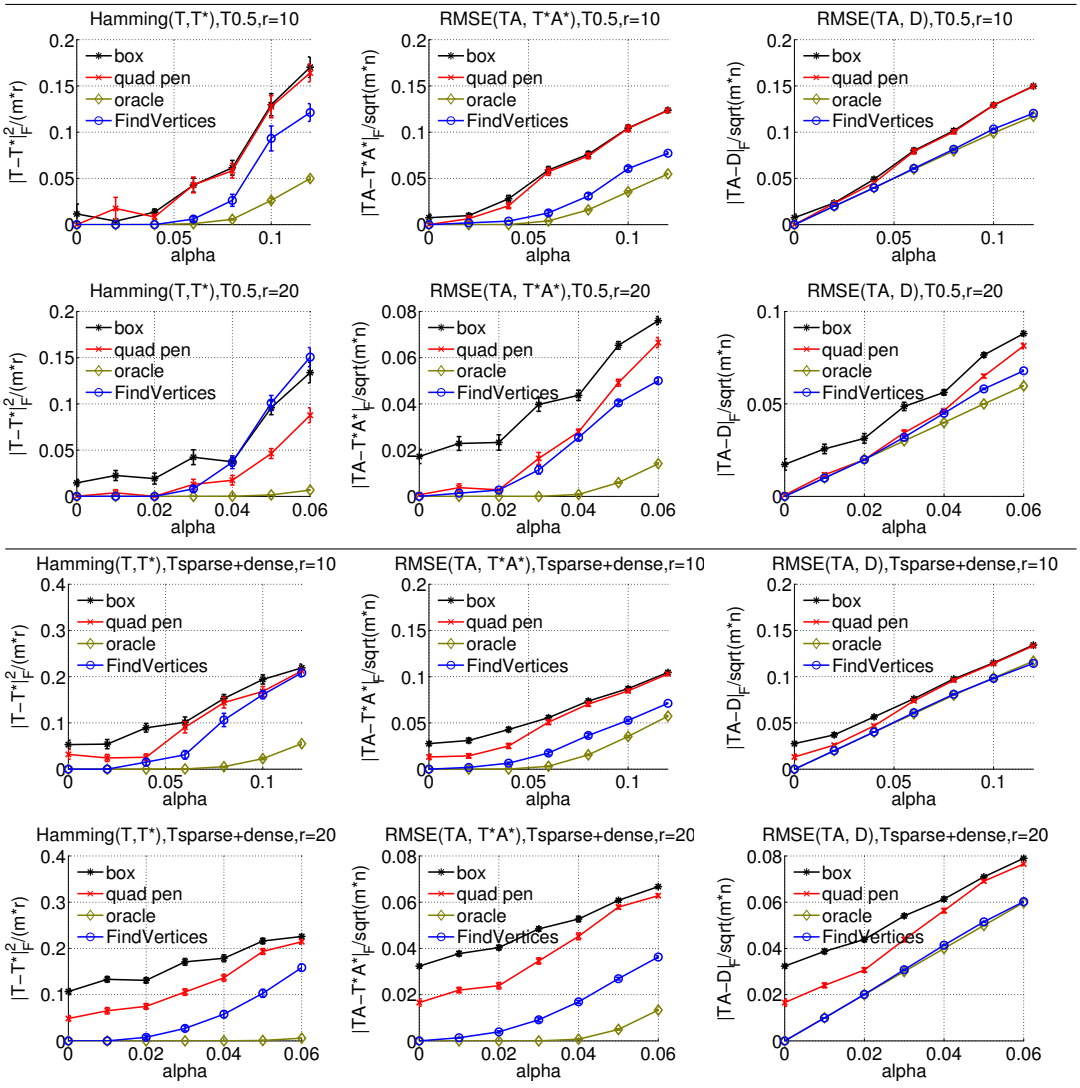


Figure 2.5: Results of the synthetic data experiments separated according to the two setups 'T.05' and 'Tsparse+dense'. Bottom/top: $r = 10$, $r = 20$. Left/Middle/Right: $\|T^* - T\|_F^2/(mr)$, $\|T^*A^* - TA\|_F/(mn)^{1/2}$ and $\|TA - D\|_F/(mn)^{1/2}$.

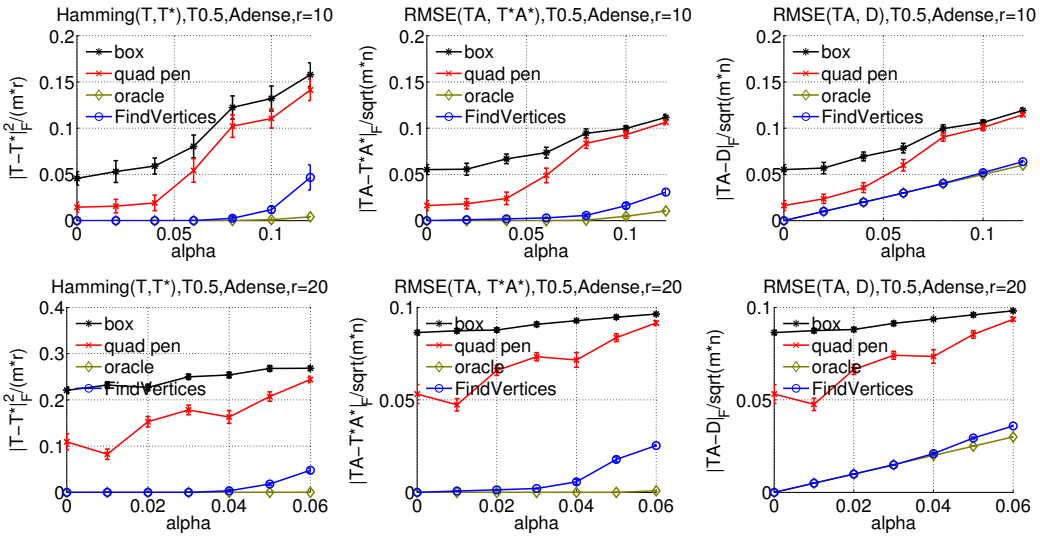


Figure 2.6: Results of the synthetic data experiments (continued) for the setup 'T05,Adense'. Bottom/top: $r = 10$, $r = 20$. Left/Middle/Right: $\|T^* - T\|_F^2/(m\,r)$, $\|T^*A^* - TA\|_F/(m\,n)^{1/2}$ and $\|TA - D\|_F/(m\,n)^{1/2}$.

Evaluation.
For each of the three setups above 20 replications are considered. We report the average performance over these replications with regard to the following criteria: the normalized Hamming distance $\|T^* - T\|_F^2/(m\,r)$ and the two RMSEs $\|T^*A^* - TA\|_F/(m\,n)^{1/2}$ and $\|TA - D\|_F/(m\,n)^{1/2}$, where (T, A) denotes the output of one of the above approaches to be compared.
From Figures 2.5 and 2.6, we find that our approach outperforms box and quad pen in most cases. At least for small levels of noise, FindVertices comes close to the oracle throughout all setups. This is unlike box and quad pen, whose performance is competitive only for the first setup. For the other two setups, not even the exact case ($\alpha = 0$) is always tackled successfully by these two methods.

Comparison to HOTTOPIXX.
According to our experimental setup, the factor A^* is non-negative, hence the problem under consideration is a special NMF problem. As mentioned in §2.3, there exist practical as well as theoretically founded algorithms for a subclass termed separable NMF problems. Since box and quad pen are heuristic approaches, it makes sense to complement the experimental comparison by assessing the performance of our approach relative to a second approach equipped with theoretical guarantees. We therefore conduct a second series of synthetic data experiments devoted to a comparison with HOTTOPIXX (HT, [13]). Since separability is crucial to the performance of HT, we restrict our comparison to separable $T = [I_r; M]$, generating the entries of M i.i.d. from a Bernoulli distribution with parameter 0.5. For runtime reasons, we lower the dimension to $m = 100$. Apart from that, the experimental setup is as for 'T0.5' above. We use a CPLEX implementation of HT available from [67]. When running that implementation, we first pre-normalize D to have unit row sums as required, and obtain A as

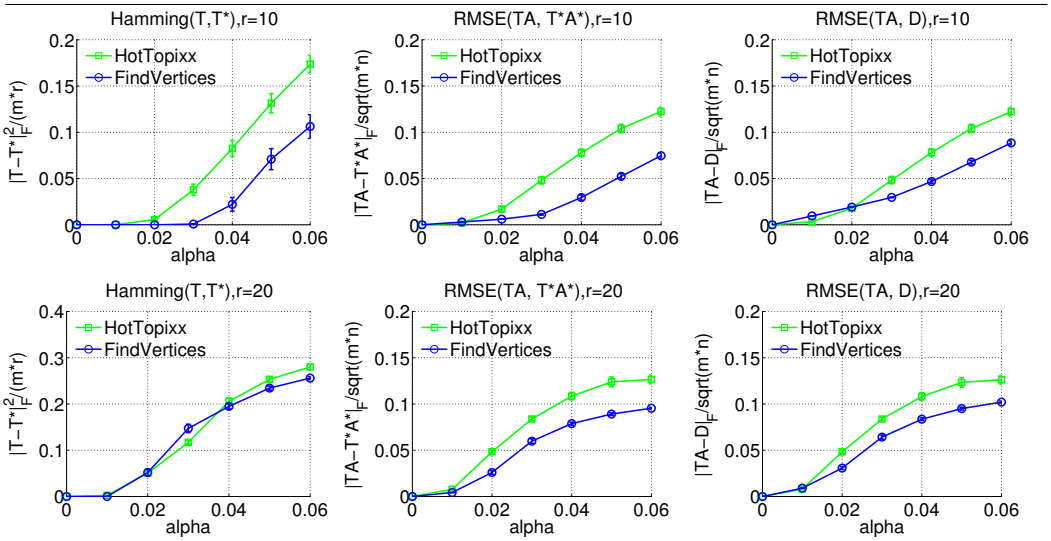


Figure 2.7: Results of the experimental comparison against HOTTOPIXX.

first output. Given A, the non-negative least squares problem $\min_{T \in \mathbb{R}_+^{m \times r}} \|D - TA\|_F^2$ is solved. Subsequently, the entries of T are re-scaled to match the original scale of D, and thresholding at 0.5 is applied to obtain a binary matrix. Finally, A is re-optimized by solving the above fitting problem with respect to A in place of T. In the noisy case, HT needs a tuning parameter to be specified that depends on the noise level, and we consider a grid of 12 values for that parameter. The range of the grid is chosen based on knowledge of the underlying additive noise matrix E. For each run, we pick the parameter that yields the best performance in favour of HT.
Figure 2.7 indicates that in the separable case, our approach performs favourably as compared to HT, a natural benchmark in this setting.

Analysis of DNA methylation data.

Background.
DNA methylation is a common chemical modification occurring at specific sites of the DNA, so-called CpGs. DNA methylation may influence cellular processes in various ways and is known to play an important role in the pathogenesis of various diseases, notably cancer [169]. Quantifying DNA methylation and relating it to phenotypes of interest (e.g. diseased and non-diseased) is regarded as a key challenge in epigenetics, a field in the life sciences concerned with changes in gene activity not caused by changes of the DNA sequence itself. DNA methylation microarrays have enabled researchers to measure methylation at a large number (up to 480k) of CpGs simultaneously.

Model.
At the level of a single cell, DNA methylation profiles are ternary, with each CpG site being methylated (1), unmethylated (0) or half-methylated (0.5); since the fraction of half-methylated sites is comparatively small, we confine ourselves to a binary

model. DNA methylation profiles as measured by microarrays involve assays containing thousands to millions of cells, each of which may have its own profile. Consequently, the resulting measurements are averages over numerous cells and take values in $[0, 1]$. However, it is assumed that the cell populations consist of few homogeneous subpopulations having a common methylation profile, which are also shared across different samples. These subpopulations are referred to as cell types (see below for a specific example). Accordingly, the measurements can be modelled as mixtures of cell type-specific profiles, where the mixture proportions differ from sample to sample. Letting $D \in [0,1]^{m \times n}$ denote the matrix of methylation profiles for m CpGs and n samples, we suppose that $D \approx TA$, where $T_{:,k} \in \{0,1\}^m$ represents the methylation profile of the k-th cell type and A_{ki} equals the proportion, i.e. $A_{ki} \geq 0$ and $\sum_{k=1}^{r} A_{ki} = 1$, of the k-th cell type in the i-th sample, $k = 1, \ldots, r$, $i = 1, \ldots, n$. The total number of cell types r is sometimes known; otherwise, it has to be determined in a data-driven way. Decomposing D in the above manner constitutes an important preliminary step before studying possible associations between phenotype and methylation, since one needs to adjust for heterogeneity of the samples w.r.t. their cell type composition. If one of T or A is (approximately) known, the missing factor can be determined by solving a constrained least squares problem. While it may be possible to obtain T and/or A via experimental techniques if sufficient prior knowledge about the composition of the cell populations is available, this tends to require considerable effort, and it is thus desirable to recover both T and A. At this point, the proposed matrix factorization comes into play.

Data set.
We consider the data set studied in [77], with $m = 500$ pre-selected CpG sites and $n = 12$ samples of blood cells composed of four major cell types (B-/T-cells, granulocytes, monocytes), i.e. $r = 4$. Ground truth is partially available: the cell type proportions of the samples, denoted by A^*, are known.

Analysis.
We apply our approach to compute an approximate factorization $D \approx \overline{T}\ \overline{A}$, $\overline{T} \in \{0.1, 0.9\}^{m \times r}$, $\overline{A} \in \mathbb{R}_+^{r \times n}$ and $\overline{A}^\top \mathbf{1}_r = \mathbf{1}_n$. We work with $\{0.1, 0.9\}$ instead of $\{0, 1\}$ in order to account for measurement noise in D that slightly pushes values towards 0.5. We first obtain $\overline{T}$ according to Algorithm 2.7 with an appropriate modification of the matrix $B^{(r-1)}$ in step 3. Given $\overline{T}$, we solve the quadratic program $\overline{A} = \operatorname{argmin}_{A \in \mathbb{R}_+^{r \times n}, A^\top \mathbf{1}_r = \mathbf{1}_n} \|D - \overline{T}A\|_F^2$ and compare $\overline{A}$ to the ground truth A^*. In order to judge the fit as well as the matrix $\overline{T}$ returned by our method, we compute $T^* = \operatorname{argmin}_{T \in \{0.1, 0.9\}^{m \times r}} \|D - TA^*\|_F^2$ as in (2.28). We obtain 0.025 as average mean squared difference of $\overline{T}$ and T^*, which corresponds to an agreement of 96 percent. Figure 2.8 indicates at least a qualitative agreement of A^* and $\overline{A}$. In the rightmost plot, we compare the RMSEs of our approach for different choices of r relative to the RMSE of (T^*, A^*). The error curve flattens after $r = 4$, which suggests that with our approach, we can recover the correct number of cell types.

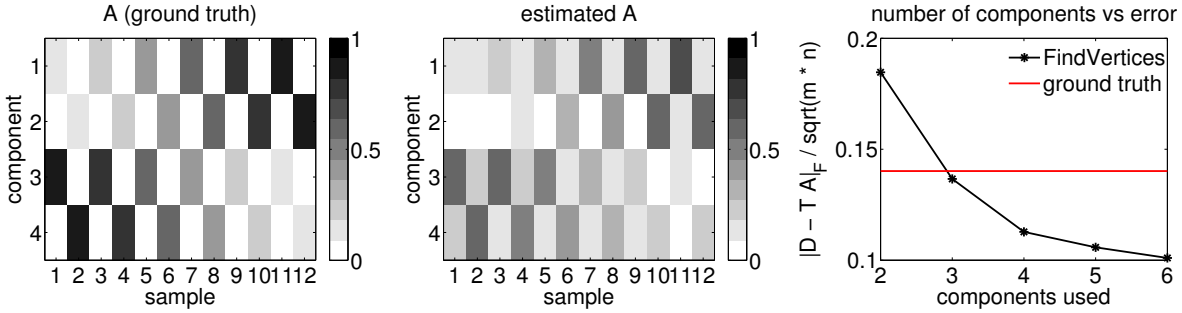


Figure 2.8: Left: Mixture proportions of the ground truth. Middle: mixture proportions as estimated by our method. Right: RMSEs $\|D - \overline{T}\ \overline{A}\|_F/(m\,n)^{1/2}$ in dependency of r.

2.4.9 Open problems

The present work on matrix factorization with binary components leaves open a number of questions pointing to interesting direction of future research. First, it is worthwhile to make the empirical observations about uniqueness and the L-O lemma rigorous in some way. While a focus has been on random binary matrices, it would also be helpful to find deterministic conditions ensuring uniqueness that apply more broadly than separability (Proposition 2.6).
Second, we do not provide any analysis for the noisy case realistically encountered in applications. Proving recovery of the binary matrix as observed empirically (cf. the left panels of Figures 2.5 and 2.6) seems to require a lower bound on

$$\operatorname{dist}(\operatorname{aff}(T), \{0,1\}^m \setminus \{T_{:,1}, \ldots, T_{:,r}\})$$

as a natural generalization of the uniqueness condition (2.19). So far, we do not have a good grasp about how such lower bound scales with m and r, which we consider as essential for a meaningful analysis. Apart from that, we expect that existing perturbation theory for the SVD can be employed to obtain a suitable result.

Appendix A

Addenda Chapter 1

A.1. Empirical scaling of $\tau^2(S)$ for $\mathbf{Ens}_+$

In §1.4.6, we have empirically investigated the scaling of $\tau^2(S)$ for the class (1.43) in a high-dimensional setting for the following designs.

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

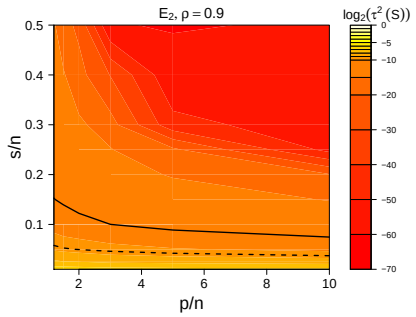
$$E_2\colon\ \{x_{ij}\} \overset{\text{i.i.d.}}{\sim}\ \tfrac{1}{\sqrt{\pi}}\,\text{Bernoulli}(\pi),\ \pi \in \{\tfrac{1}{10}, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4}, \tfrac{9}{10}\}\ \ (\rho \in \{\tfrac{1}{10}, \tfrac{1}{4}, \tfrac{1}{2}, \tfrac{3}{4}, \tfrac{9}{10}\})$$

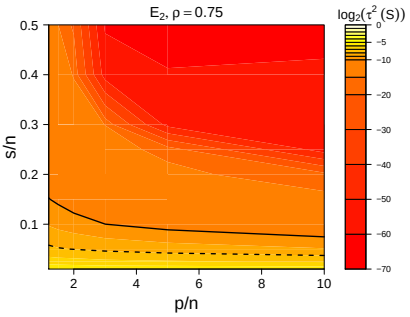
$$E_3\colon\ \{x_{ij}\} \overset{\text{i.i.d.}}{\sim}\ |Z|,\ Z \sim a\,\text{Gaussian}(0,1) + (1-a)\delta_0,\ a \in \{1, \tfrac{\pi}{4}, \tfrac{\pi}{8}, \tfrac{\pi}{20}\}\ \ (\rho \in \{\tfrac{2}{\pi}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{10}\})$$

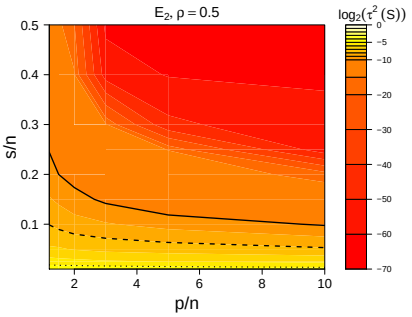
$$E_4\colon\ \{x_{ij}\} \overset{\text{i.i.d.}}{\sim}\ a\text{Poisson}(3/\sqrt{12\,a}) + (1-a)\delta_0,\ a \in \{1, \tfrac{2}{3}, \tfrac{1}{3}, \tfrac{2}{15}\}\ \ (\rho \in \{\tfrac{3}{4}, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{10}\})$$

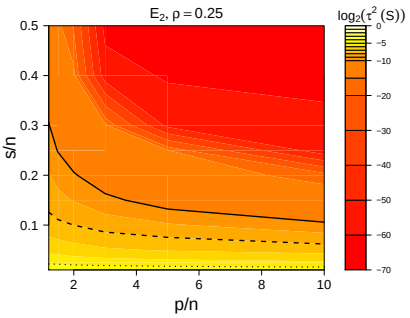
The results for E_1 are displayed in Figure 1.4, and the results for E_2 to E_4 are displayed below.

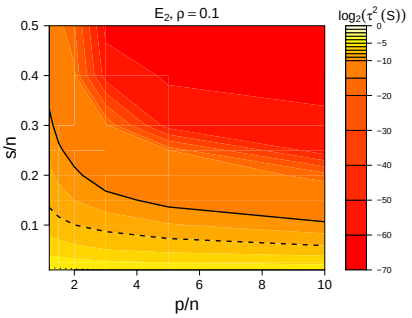
E_2



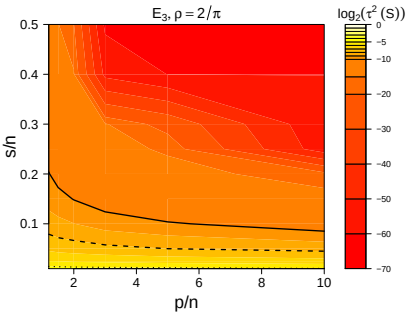

E2, ρ = 0.75
log2(τ2(S))
s/n
p/n

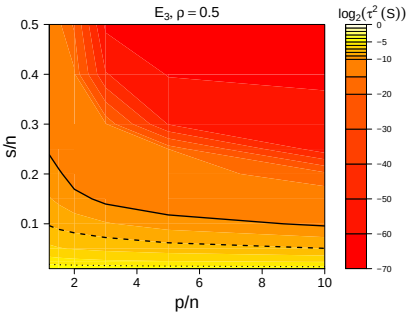

E2, ρ = 0.5
log2(τ2(S))
s/n
p/n

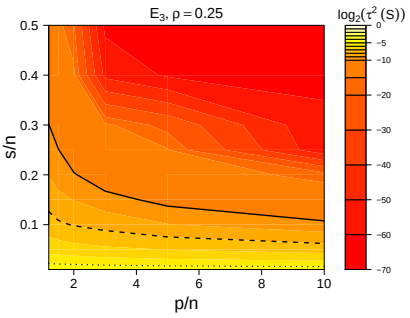

E2, ρ = 0.25
log2(τ2(S))
s/n
p/n

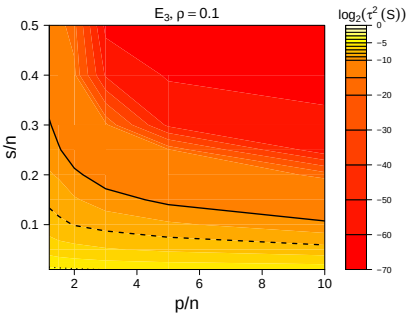

E2, ρ = 0.1
log2(τ2(S))
s/n
p/n

E_3

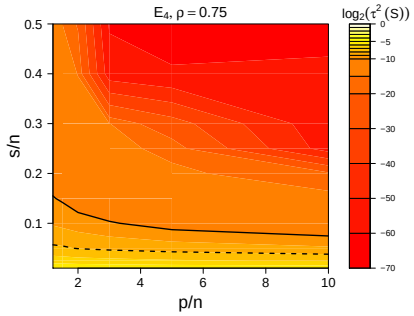

E3, ρ = 2/π
log2(τ2(S))
s/n
p/n

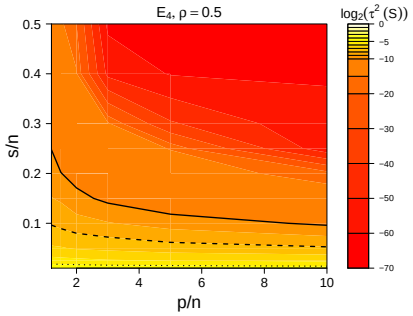

E3, ρ = 0.5
log2(τ2(S))
s/n
p/n

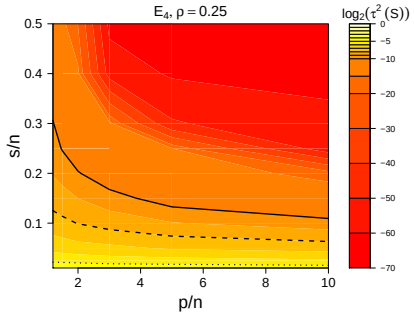

E3, ρ = 0.25
log2(τ2(S))
s/n
p/n

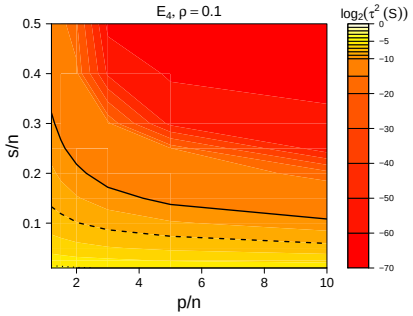

E3, ρ = 0.1
log2(τ2(S))
s/n
p/n

E_4


E4, ρ = 0.75
log2(τ2 0(S))
s/n
p/n


E4, ρ = 0.5
log2(τ2 0(S))
s/n
p/n


E4, ρ = 0.25
log2(τ2 0(S))
s/n
p/n


E4, ρ = 0.1
log2(τ2 0(S))
s/n
p/n

Appendix B

Addenda Chapter 2

B.1. Example of non-uniqueness under non-negativity of the right factor

We here show that even the non-negative variant (2.20) may not have a unique solution. For this purpose, we construct $T, \widetilde{T} \in \{0,1\}^{m\times r}$ and $A, \widetilde{A} \in \mathbb{R}_+^{r\times n}$, $A^\top \mathbf{1}_r = \widetilde{A}^\top \mathbf{1}_r = \mathbf{1}_n$ so that $TA = \widetilde{T}\widetilde{A}$ while $\widetilde{T}$ is not a column permutation of T. For $r > 2$ and $k = r-1$, consider

$$T = \begin{pmatrix} & 0_{(m-k)\times r} & \\ I_{k-1} & \mathbf{1}_{k-1} & 0_{k-1} \\ 0_{k-1}^\top & 1 & 0 \end{pmatrix},$$

Observe that the columns of T are affinely independent. Let further $A = [A^{(1)}, A^{(2)}]$ consist of two blocks $A^{(1)}$ and $A^{(2)}$ such that

$$A^{(1)} = [\Gamma;\, 0_{n_1}^\top], \qquad A^{(2)} = [0_{(r-2)\times n_1};\, \alpha^\top;\, (\mathbf{1}_{n_2} - \alpha)^\top],$$

where $n_1 + n_2 = n$, $\Gamma \in \mathbb{R}_+^{(r-1)\times n_1}$, and $\alpha \in (0,1)^{n_2}$.
Consider now

$$\widetilde{T} = \begin{pmatrix} & 0_{(m-k)\times r} & \\ I_{k-1} & \mathbf{1}_{k-1} & 0_{k-1} \\ 0_{k-1}^\top & 1 & 1 \end{pmatrix}$$

and $\widetilde{A} = [\widetilde{A}^{(1)}, \widetilde{A}^{(2)}]$ with $\widetilde{A}^{(1)} = A^{(1)}$ and

$$\widetilde{A}^{(2)} = \begin{pmatrix} \gamma_1 \mathbf{1}_{r-2} & \dots & \gamma_{n_2} \mathbf{1}_{r-2} \\ \eta_1 & \dots & \eta_{n_2} \\ \gamma_1 & \dots & \gamma_{n_2} \end{pmatrix}$$

where (γ_i, η_i) satisfy

$$\gamma_i, \eta_i > 0, \qquad \alpha_i = \gamma_i + \eta_i, \qquad (r-1)\gamma_i + \eta_i = 1, \;\; i = 1, \dots, n_2. \tag{B.1}$$

It then holds that $TA = \widetilde{T}\widetilde{A}$. This can be verified as follows. First note that because of $T_{:,1:(r-1)} = \widetilde{T}_{:,1:(r-1)}$ and the fact that the r-th row of $A^{(1)} = \widetilde{A}^{(1)}$ equals zero, it holds that $TA^{(1)} = \widetilde{T}\widetilde{A}^{(1)}$. It thus remains to verify that $TA^{(2)} = \widetilde{T}\widetilde{A}^{(2)}$. We have

$$TA^{(2)} = T_{:,r-1}\alpha^\top + T_{:,r}(1-\alpha)^\top = \begin{pmatrix} 0_{(m-k)\times n_2} \\ \mathbf{1}_k \alpha^\top \end{pmatrix} \tag{B.2}$$

On the other hand, with $\gamma = (\gamma_1, \ldots, \gamma_{n_2})^\top$ and $\eta = (\eta_1, \ldots, \eta_{n_2})^\top$, we have

$$\begin{aligned}\widetilde{T}\widetilde{A}^{(2)} &= \widetilde{T}_{:,1:(r-2)}\mathbf{1}_{r-2}\gamma^\top + \widetilde{T}_{:,r-1}\eta^\top + \widetilde{T}_{:r}\gamma^\top \\ &= \begin{pmatrix} 0_{(m-k)\times n_2} \\ \mathbf{1}_{k-1}\gamma^\top \\ 0_{n_2}^\top \end{pmatrix} + \begin{pmatrix} 0_{(m-k)\times n_2} \\ \mathbf{1}_{k-1}\eta^\top \\ \eta^\top \end{pmatrix} + \begin{pmatrix} 0_{(m-k)\times n_2} \\ \mathbf{0}_{(k-1)\times n_2} \\ \gamma^\top \end{pmatrix} = \begin{pmatrix} 0_{(m-k)\times n_2} \\ \mathbf{1}_k(\gamma+\eta)^\top \end{pmatrix}\end{aligned}$$

The claim now follows from (B.1) and (B.2).

B.2. Optimization for the quadratic penalty-based approach

We here discuss in a bit of more detail our approach to the optimization problem (2.29)

$$\min_{\substack{T\in[0,1]^{m\times r} \\ A\in\mathbb{R}_+^{r\times n},\, A^\top\mathbf{1}_r=\mathbf{1}_n}} \|D - TA\|_F^2 + \lambda\sum_{i=1}^{m}\sum_{k=1}^{r}\omega(T_{ik}), \quad \text{where } \omega(t) = t(1-t).$$

At a basic level, we use block coordinate descent in which T and A are optimized in an alternating fashion, cf. Algorithm 2.1. The update for T involves minimization of a function which is the sum of one convex and one concave function, i.e.

$$\min_{T\in[0,1]^{m\times r}} g(T) + h(T), \quad \text{where } g(T) = \|D - TA\|_F^2, \quad \text{and } h(T) = \lambda\sum_{i=1}^{m}\sum_{k=1}^{r}\omega(T_{ik}). \tag{B.3}$$

We thus make use of a technique known as convex-concave procedure (CCCP, [172]) tailored to this specific structure. The main idea is to tightly upper bound the objective function at the current iterate by linearizing the concave part. The result is again a convex quadratic function, whose minimizer is used as next iterate. One then alternates between these two steps. As shown below, this approach ensures that the objective function decreases at each iteration. Since ω is concave, $-\omega$ is convex. From the first order convexity condition (e.g. [16], p. 70), we have

$$\omega(y) \le \omega(x) + \omega'(x)(y-x) \quad \forall x, y \in \mathbb{R}.$$

Letting T^t denote the current iterate of the variable T in (B.3), we define a matrix G having entries $G_{ik} = \lambda\omega'(T_{ik}^t),\ i = 1, \ldots, m,\ k = 1, \ldots, r$, and summing up above inequality over all entries, we obtain that for any T

$$\begin{aligned}h(T) = \lambda\sum_{i=1}^{m}\sum_{k=1}^{r}\omega(T_{ik}) &\le \lambda\left\{\sum_{i=1}^{m}\sum_{k=1}^{r}\omega(T_{ik}^t) + \sum_{i=1}^{m}\sum_{k=1}^{k}\omega'(T_{ik}^t)(T_{ik} - T_{ik}^t)\right\} \\ &= h(T^t) + \operatorname{tr}((T - T^t)^\top G).\end{aligned}$$

It follows immediately that for any T

$$g(T) + h(T) \le g(T) + h(T^t) + \operatorname{tr}((T - T^t)^\top G) \tag{B.4}$$

The right hand side tightly upper bounds the objective function at $T = T^t$. This observation implies that

$$\min_T g(T) + h(T^t) + \mathrm{tr}((T - T^t)^\top G) \leq g(T^t) + h(T^t), \tag{B.5}$$

i.e. minimizing the right hand side of (B.4) yields descent for the original objective function (B.3).

Bibliography

[1] R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Restricted Isometry Property of matrices with independent columns and neighborly polytopes by random sampling. *Constructive Approximation*, 34:61–88, 2011.

[2] E. Amaldi and V. Kann. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science*, 209:237–260, 1998.

[3] S. Arora, R. Ge, R. Kannan, and A. Moitra. Computing a nonnegative matrix factorization – provably. In *Symposium on Theory of Computing (STOC)*, pages 145–162, 2012.

[4] A. Banerjee, C. Krumpelman, J. Ghosh, S. Basu, and R. Mooney. Model-based overlapping clustering. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 532–537, 2005.

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

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

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

[8] A. Berman and R. Plemmons. *Nonnegative matrices in the mathematical sciences*. SIAM Classics in Applied Mathematics, 1994.

[9] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[10] P. Bickel, J. Brown, H. Huang, and Q. Li. An overview of recent developments in genomics and associated statistical methods. *Philosophical Transactions of the Royal Society A*, 367:4313–4337, 2009.

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

[12] E. Birgin, J. Martinez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM Journal on Optimization*, 10:1196–1211, 2000.

[13] V. Bittdorf, B. Recht, C. Re, and J. Tropp. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1214–1222. 2012.

[14] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27:265–274, 2009.

[15] T. Blumensath, M. Davies, and G. Rilling. *Compressed Sensing: Theory and Applications*, chapter 'Greedy Algorithms for Compressed Sensing'. Cambridge University Press, 2012.

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

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

[18] A. Bruckstein, D. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

[19] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer, 2011.

[20] V. Buldygin and Y. Kozachenko. Sub-Gaussian random variables. *Ukrainian Mathematical Journal*, 32:483–489, 1980.

[21] F. Bunea. 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.

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

[23] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35:1674–1697, 2007.

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

[25] T. Cai and A. Zhang. ROP: Matrix recovery via rank-one projections. arXiv:1310.5791, 2013.

[26] T. Cai, W. Liu, and X. Luo. A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

[27] E. Candes and M. Davenport. How well can we estimate a sparse vector ? *Applied and Computational Harmonic Analysis*, 34:317–323, 2013.

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

[29] E. Candes and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of noisy measurements. *IEEE Transactions on Information Theory*, 57:2342–2359, 2011.

[30] E. Candes and T. Tao. Decoding by Linear Programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[31] E. Candes and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies. *IEEE Transactions on Information Theory*, 52: 5406–5425, 2006.

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

[33] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006.

[34] D. Chafai, O. Guedon, G. Lecue, and A. Pajor. *Interactions between Compressed Sensing, Random Matrices and High-dimensional Geometry*. Lecture Notes, Université Paris Est Marne-la Vallée, 2012. http://djalil.chafai.net/Docs/Livres/LAMABOOK/lamabook-draft.pdf.

[35] D. Chen and R. Plemmons. Nonnegativity constraints in numerical analysis. In *Symposium on the Birth of Numerical Analysis*, pages 109–140, 2009.

[36] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1999.

[37] Y. Chen, Y. Chi, and A. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. arXiv:1310.0807, 2013.

[38] A. Cichoki, R. Zdunek, A. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations*. Wiley, 2009.

[39] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22:211–231, 2009.

[40] J. Dattorro. *Convex optimization and Euclidean distance geometry*. Meboo Publishing USA, 2005.

[41] C. Ding, T. Li, and M. Jordan. Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Tranasctions on Pattern Analysis and Machine Intelligence*, 32: 45–55, 2010.

[42] T. Pham Dinh and H. Le Thi. Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22:289–355, 1997.

[43] D. Donoho. Neighborly Polytopes and Sparse Solutions of Underdetermined Linear Equations. Technical report, Stanford University, 2004.

[44] D. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59:907–034, 2006.

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

[46] D. Donoho. High-Dimensional Centrosymmetric Polytopes with Neighborliness Proportional to Dimension. *Discrete and Computational Geometry*, 35:617–652, 2006.

[47] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47:2845–2862, 2001.

[48] D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems (NIPS)*, pages 1141–1148. 2003.

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

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

[51] D. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22:1–53, 2009.

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

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

[54] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936.

[55] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least Angle Regression. *The Annals of Statistics*, 32:407–499, 2004.

[56] Y. Eldar and G. Kutyniok, editors. *Compressed sensing: theory and applications*. Cambridge University Press, 2012.

[57] P. Erdös. On a lemma of Littlewood and Offord. *Bulletin of the American Mathematical Society*, 51:898–902, 1951.

[58] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 97:210–221, 2001.

[59] J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics*, 3:521–541, 3.

[60] J. Fan, F. Han, and H. Liu. Challenges of Big Data analysis. *National Science Review*, to appear.

[61] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49:1579–1581, 2003.

[62] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

[63] D. Gale. Neighborly and cyclic polytopes. In V. Klee, editor, *Symposia in Pure Mathematics*, volume 7, pages 225–232, 1963.

[64] G. Gasso, A. Rakotomamonjy, and S. Canu. Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing*, 57:4686–4698, 2009.

[65] C. Genovese, J. Jin, L. Wasserman, and Z. Yao. A Comparison of the Lasso and Marginal Regression. *Journal of Machine Learning Research*, 13:2107–2143, 2012.

[66] N. Gillis. *Nonnegative Matrix Factorization: Complexity, Algorithms and Applications*. PhD thesis, Université catholique de Louvain, 2011.

[67] N. Gillis. CPLEX implementation of HOTTOPIXX . https://sites.google.com/site/nicolasgillis/publications, 2013.

[68] N. Gillis and S. Vavasis. Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear.

[69] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

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

[71] E. Grushka. Characterization of Exponentially Modified Gaussian Peaks in Chromatography. *Analytical Chemistry*, 44:1733–1738, 1972.

[72] M. Gu and S. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17:848–869, 1996.

[73] N. Halko, P. Martinsson, and J. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53:271–288, 2011.

[74] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. 2nd edition.

[75] M. Hebiri and J. Lederer. How correlations influence Lasso Prediction. *IEEE Transactions on Information Theory*, 59:1846–1854, 2013.

[76] M. Hofmann, C. Gatu, and E. Kontoghiorghes. Efficient algorithms for computing the best subset regression models for large-scale problems. *Computational Statistics & Data Analysis*, 52:16–29, 2007.

[77] E. Houseman, W. Accomando, D. Koestler, B. Christensen, C. Marsit, H. Nelson, J. Wiencke, and K. Kelsey. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, 13:86, 2012.

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

[79] R. Hussong, A. Tholey, and A. Hildebrandt. Efficient Analysis of Mass Spectrometry Data Using the Isotope Wavelet. In *COMPLIFE 2007*, volume 940, pages 139–149, 2007.

[80] I. Johnstone. Gaussian estimation: Sequence and wavelet models. http://statweb.stanford.edu/~imj/GE06-11-13.pdf, 2013.

[81] I. Johnstone and A. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104: 682–693, 2009.

[82] I. Johnstone and D. Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A*, 367:4237–4253, 2009.

[83] A. Kaban and E. Bingham. Factorisation and denoising of 0-1 data: a variational approach. *Neurocomputing*, 71:2291–2308, 2008.

[84] J. Kahn, J. Komlos, and E. Szemeredi. On the Probability that a ± 1 matrix is singular. *Journal of the American Mathematical Society*, 8:223–240, 1995.

[85] P. Kaur and P. B. O'Connor. Algorithms for automatic interpretation of high resolution mass spectra. *Journal of the American Society for Mass Spectrometry*, 17:459–468, 2006.

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

[87] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[88] A. Kumar, V. Sindhwani, and P. Kambadur. Fast Conical Hull Algorithms for Near-separable Non-negative Matrix Factorization. In *International Conference on Machine Learning (ICML)*, pages 231–239, 2013.

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

[90] S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[91] R. Lawson and C. Hanson. *Solving least squares problems*. SIAM Classics in Applied Mathematics, 1987.

[92] D. Lee and H. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[93] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562, 2000.

[94] E. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998. 2nd edition.

[95] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16:1273–1284, 2006.

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

[97] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural Computation*, 19:2756–2779, 2007.

[98] Y. Lin, D. Lee, and L. Saul. Nonnegative deconvolution for time of arrival estimation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 377–380, 2004.

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

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

[101] D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[102] J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. In *International Conference on Machine Learning (ICML)*, pages 353–360, 2012.

[103] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[104] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[105] V. Marco and G. Bombi. Mathematical functions for the representation of chromatographic peaks. *Journal of Chromatography A*, 931:1–30, 2001.

[106] P. Massart. *Concentration inequalities and model selection*. Springer, 2007.

[107] E. Meeds, Z. Gharamani, R. Neal, and S. Roweis. Modeling dyadic data with binary latent factors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 977–984, 2007.

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

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

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

[111] M. Mendozza, M. Raydan, and P. Tarazaga. Computing the nearest diagonally dominant matrix. *Numerical Linear Algebra with Applications*, 5:461–474, 1998.

[112] C. Michelot. A finite algorithm for finding the projection of a point onto the Canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50: 195–200, 1986.

[113] A. Miller. *Subset Selection in Regression*. Chapman & Hall, 2002.

[114] B. Natarjan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24, 1995.

[115] S. Negahban and M. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39:1069–1097, 2011.

[116] H. Nguyen and V. Vu. Small ball probability, inverse theorems, and applications. arXiv:1301.0019.

[117] A. Odlyzko. On Subspaces Spanned by Random Selections of ± 1 vectors. *Journal of Combinatorial Theory A*, 47:124–133, 1988.

[118] P. Miettinen and T. Mielikäinen and A. Gionis and G. Das and H. Mannila. The discrete basis problem. In *European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 335–346, 2006.

[119] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[120] D. Perkins and D. Pappin. MASCOT search engine. www.matrixscience.com, 2012. Version 2.2.04.

[121] J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, New York, 2006.

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

[123] G. Raskutti, M. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear models over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57:1069–1097, 2011.

[124] H. Rauhut and S. Foucart. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[125] P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *The Electronic Journal of Statistics*, 4:935–980, 2011.

[126] M. Raydan and P. Tarazaga. Primal and polar approaches for computing the symmetric diagonally dominant projection. *Numerical Linear Algebra with Applications*, 9:333–345, 2002.

[127] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52:471–501, 2010.

[128] B. Renard, M. Kirchner, H. Steen, J. Steen, and F. Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics*, 9:355, 2008.

[129] T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970. reprint 1997.

[130] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 1012-1030:35, 2007.

[131] A. Rothman, P. Bickel, L. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *The Electronic Journal of Statistics*, 2:494–515, 2008.

[132] M. Rudelson and R. Vershynin. Geometric approach to error correcting codes and reconstruction of signals. *International Mathematical Research Notices*, 64: 4019–4041, 2005.

[133] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *International Congress of Mathematics*, pages 1576–1602, 2010.

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

[135] J. Samuelsson, D. Dalevi, F. Levander, and T. Rognvaldsson. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics*, 20:3628–3635, 2004.

[136] A. Schein, L. Saul, and L. Ungar. A generalized linear model for principal component analysis of binary data. In *International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2003.

[137] M. Schmidt, G. Fung, and R. Rosales. Fast Optimization Methods for L1-Regularization: A Comparative Study and Two New Approaches. In *European Conference on Machine Learning (ECML)*, pages 286–297, 2007.

[138] B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.

[139] O. Schulz-Trieglaff. *Computational Methods for Quantitative Peptide Mass Spectrometry*. PhD thesis, Freie Universität Berlin, 2008.

[140] O. Schulz-Trieglaff, R. Hussong, C. Gröpl, A. Hildebrandt, and K. Reinert. A Fast and Accurate Algorithm for the Quantification of Peptides from Mass Spectrometry Data. In *Proceedings of the Eleventh Annual International Conference on Research in Computational Molecular Biology (RECOMB 2007)*, pages 437–487, 2007.

[141] E. Segal, A. Battle, and D. Koller. Decomposing gene expression into cellular processes. In *Proceedings of the 8th Pacific Symposium on Biocomputing*, pages 89–100, 2003.

[142] M. Senko, S. Beu, and F. McLafferty. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *Journal of the American Society for Mass Spectrometry*, 6:229–233, 1995.

[143] M. Slawski and M. Hein. Positive definite M-matrices and structure learning in attractive Gaussian Markov random fields. *Linear Algebra and its Applications*, in press.

[144] J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse Image and Signal Processing: wavelets, curvelets, morphological diversity.* Cambridge University Press, 2010.

[145] F. Suits, B. Hoekman, T. Rosenling, R. Bischoff, and P. Horvatovich. Threshold-avoiding proteomics pipeline. *Analytical Chemistry*, 83:7786–7794, 2011.

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

[147] A. Szlam, Z. Guo, and S. Osher. A split Bregman method for non-negative sparsity penalized least squares with applications to hyperspectral demixing. In *International Conference on Image Processing*, pages 1917–1920, 2010.

[148] T. Tao and V. Vu. On the singularity problem of random Bernoulli matrices. *Journal of the American Mathematical Society*, 20:603–628, 2007.

[149] T. Tao and V. Vu. The Littlewoord-Offord problem in high-dimensions and a conjecture of Frankl and Füredi. *Combinatorica*, 32:363–372, 2012.

[150] M. Thiao. *Approches de la programmation DC et DCA en data mining. Modélisation parcimonieuse de données.* PhD thesis, INSA Rouen, 2011.

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

[152] A. Tillmann and M. Pfetsch. The Computational Complexity of RIP, NSP, and Related Concepts in Compressed Sensing. *IEEE Transactions on Information Theory*, 60:1248–1259, 2014.

[153] A. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2008.

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

[155] S. Tu, R. Chen, and L. Xu. Transcription Network Analysis by a Sparse Binary Factor Analysis Algorithm. *Journal of Integrative Bioinformatics*, 9:198, 2012.

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

[157] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *The Electronic Journal of Statistics*, 3:1360–1392, 2009.

[158] S. van de Geer and J. Lederer. *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.

[159] A-J. van der Veen. Analytical Method for Blind Binary Signal Separation. *IEEE Transactions on Signal Processing*, 45:1078–1082, 1997.

[160] S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20:1364–1377, 2009.

[161] R. Vershynin. *In: Compressed Sensing: Theory and Applications, Y. Eldar and G. Kutyniok (eds.)*, chapter 'Introduction to the non-asymptotic analysis of random matrices'. Cambridge University Press, 2012.

[162] R. Vidal, Y. Ma, and S. Sastry. *Generalized Principal Component Analysis*. Springer, 2014.

[163] M. Wainwright. 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.

[164] L. Wang, Y. Kim, and R. Li. Calibrating nonconvex penalized regression in ultra-high dimension. *The Annals of Statistics*, 41:2263–2702, 2013.

[165] M. Wang and A. Tang. Conditions for a Unique Non-negative Solution to an Underdetermined System. In *Allerton Conference on Communication, Control, and Computing*, pages 301–307, 2009.

[166] M. Wang, W. Xu, and A. Tang. A unique nonnegative solution to an undetermined system: from vectors to matrices. *IEEE Transactions on Signal Processing*, 59:1007–1016, 2011.

[167] S. Weisberg. *Applied Linear Regression*. Wiley, 1980.

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

[169] Wikipedia. DNA methylation. http://en.wikipedia.org/wiki/DNA_methylation, 2014.

[170] L. Wolsey. *Integer Programming*. Wiley, 1998.

[171] F. Ye and C.-H. Zhang. Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, 11:3519–3540, 2010.

[172] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.

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

[174] C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27:576–593, 2013.

[175] T. Zhang. On the Consistency of Feature Selection using Greedy Least Squares Regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

[176] T. Zhang. Some Sharp Performance Bounds for Least Squares Regression with L_1 Regularization. *The Annals of Statistics*, 37:2109–2144, 2009.

[177] T. Zhang. Adaptive Forward-Backward Greedy Algorithm for Learning Sparse Representations. *IEEE Transactions on Information Theory*, 57:4689–4708, 2011.

[178] Y. Zhang. A simple proof for recoverability of ℓ_1-minimization: go over or under? Technical report, Rice University, 2005.

[179] Z. Zhang, C. Ding, T. Li, and X. Zhang. Binary matrix factorization with applications. In *International Conference on Data Mining (ICDM)*, pages 391–400, 2007.

[180] P. Zhao and B. Yu. On model selection consistency of the lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.

[181] S. Zhou. Thresholding procedures for high dimensional variable selection and statistical estimation. In *Advances in Neural Information Processing (NIPS)*, pages 2304–2312, 2009.

[182] S. Zhou. Restricted Eigenvalue Conditions on Subgaussian Random Matrices. Technical report, Department of Statistics, University of Michigan, 2011.

[183] G. Ziegler. *Lectures on polytopes*. Graduate Texts in Mathematics. Springer, 1995. Updated 7th edition of first priting.

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

[185] R. Zubarev. Accurate Monoisotopic Mass Measurements of Peptides: Possibilities and Limitations of High Resolution Time-of-flight Particle Desorption Mass Spectrometry. *Rapid Communications in Mass Spectrometry*, 10:1386–1392, 1996.