

## METHODS FOR SPARSE AND LOW-RANK RECOVERY UNDER SIMPLEX CONSTRAINTS

Ping Li, Syama Sundar Rangapuram and Martin Slawski

*Baidu Research, Amazon Research and George Mason University*

*Abstract:* The de facto standard approach of promoting sparsity by means of  $\ell_1$ -regularization becomes ineffective in the presence of simplex constraints, that is, when the target is known to have non-negative entries summing to a given constant. The situation is analogous for the use of nuclear norm regularization for the low-rank recovery of Hermitian positive semidefinite matrices with a given trace. In the present paper, we discuss several strategies to deal with this situation, from simple to more complex. First, we consider empirical risk minimization (ERM), which has similar theoretical properties w.r.t. prediction and  $\ell_2$ -estimation error as  $\ell_1$ -regularization. In light of this, we argue that ERM combined with a subsequent sparsification step (e.g., thresholding) represents a sound alternative to the heuristic of using  $\ell_1$ -regularization after dropping the sum constraint and the subsequent normalization. Next, we show that any sparsity-promoting regularizer under simplex constraints cannot be convex. A novel sparsity-promoting regularization scheme based on the inverse or negative of the squared  $\ell_2$ -norm is proposed, which avoids the shortcomings of various alternative methods from the literature. Our approach naturally extends to Hermitian positive semidefinite matrices with a given trace.

*Key words and phrases:* D.C. programming, density matrices of quantum systems, estimation of mixture proportions, simplex constraints, sparsity.

### 1. Introduction

In this paper, we study the case in which the parameter of interest  $\beta^*$  is sparse and non-negative with a known sum, i.e.,  $\beta^* \in c\Delta^p \cap \mathbb{B}_0^p(s)$ , where, for  $c > 0$  and  $1 \leq s \leq p$ ,  $c\Delta^p = \{\beta \in \mathbb{R}_+^p : \mathbf{1}^\top \beta = c\}$  is the (scaled) canonical simplex in  $\mathbb{R}^p$ ,  $\mathbb{B}_0^p(s) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s\}$ , and  $\|\beta\|_0 = |S(\beta)| = |\{j : \beta_j \neq 0\}|$  is referred to as the  $\ell_0$ -norm (the cardinality of the support  $S(\beta)$ ). Unlike the constant  $c$ , the sparsity level  $s$  is regarded as unknown. The specific value of  $c$  is not essential; in the sequel, we shall work with  $c = 1$ , as for all problem instances studied herein, the data can be re-scaled accordingly. The elements of  $\Delta^p = \{\beta \in \mathbb{R}_+^p : \mathbf{1}^\top \beta = 1\}$  can represent probability distributions over  $p$  items, proportions, or normalized weights. The following are examples of quantities

that arise frequently in contemporary data analyses:

- *Estimation of proportions.* Specific examples include determining the proportions of chemical constituents in a given sample and endmember composition of pixels in hyperspectral imaging (Keshava (2003)).
- *Probability density estimation,* cf. Bunea et al. (2010). Let  $(\mathcal{Z}, \mathcal{A}, P)$  be a probability space, with  $P$  having a density  $f$  w.r.t. some dominating measure  $\nu$ . Given a sample  $\{Z_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P$  and a dictionary  $\{\phi_j\}_{j=1}^p$  of densities (w.r.t.  $\nu$ ), the goal is to find a mixture density  $\phi_\beta = \sum_{j=1}^p \beta_j \phi_j$  that well approximates  $f$ , where  $\beta \in \Delta^p$ .
- *Convex aggregation/ensemble learning.* The following problem has attracted much interest in the field of non-parametric estimation; see Nemirovski (2000). Let  $f$  be the target in a regression/classification problem, and let  $\{\phi_j\}_{j=1}^p$  be an ensemble of regressors/classifiers. The goal is to approximate  $f$  by a convex combination of  $\{\phi_j\}_{j=1}^p$ .
- *Markowitz portfolios* (Markowitz (1952)) *without short positions.* Given assets with expected returns  $r = (r_j)_{j=1}^p$  and covariance  $\Sigma$ , the goal is to invest according to proportions  $\beta \in \Delta^p$  s.t. the variance  $\beta^\top \Sigma \beta$  is minimized, subject to a lower bound on the expected return  $\beta^\top r$ .

Sparsity is often prevalent or desired in these applications.

- In hyperspectral imaging, a single pixel usually contains few endmembers.
- In density estimation, the underlying density may be concentrated in certain regions of the sample space.
- In aggregation, it is common to work with a large ensemble to improve the approximation capacity, although specific functions may be well approximated by just a few members of the ensemble.
- Portfolios involving only few assets incur less transaction costs and are easier to manage.

At the same time, promoting sparsity in the presence of the constraint  $\beta \in \Delta^p$  appears to be more difficult, as  $\ell_1$ -regularization no longer serves this purpose. As clarified in §2, the naive approach of employing  $\ell_1$ -regularization and dropping the sum constraint results in discarded information. The situation is similar for nuclear norm regularization and low-rank matrices that are Hermitian

positive semidefinite, with a fixed trace. For example, this arises in quantum state tomography (Gross et al. (2010)) when the constraint set results as  $\Delta^m = \{B \in \mathbb{C}^{m \times m} : B = B^H, B \succeq 0, \text{tr}(B) = 1\}$ , with  $^H$  denoting conjugate transposition. Thus, the presence of simplex constraints and their matrix counterparts require that we use different strategies to deal with sparsity and low-rankedness. Here, we propose strategies that are statistically sound, straightforward to implement, adaptive, in the sense that the sparsity level  $s$  (resp., the rank in the matrix case) is not required to be known, and work with a minimum amount of hyperparameter tuning.

**Related work.** The problem outlined above is discussed well by Kyriallidis et al. (2013). They consider the sparsity level  $s$  to be known, and suggest dealing with the constraint set  $\Delta_0^p(s) = \Delta^p \cap \mathbb{B}_0^p(s)$  by projected gradient descent based on a near-linear time algorithm used to compute the projection. This approach can be viewed as a natural extension of iterative hard thresholding (IHT, Blumensath and Davies (2009); Shen and Li (2018)).

Pilanci, Ghaoui and Chandrasekaran (2012) suggest using the regularizer  $\beta \mapsto 1/\|\beta\|_\infty$  to promote sparsity on  $\Delta^p$ . In addition, they show that in the case of squared loss, the resulting nonconvex optimization problem can be reduced to  $p$  second-order cone programs. In practice, however, the computational cost quickly becomes prohibitive, particularly when combined with the tuning of the regularization parameter.

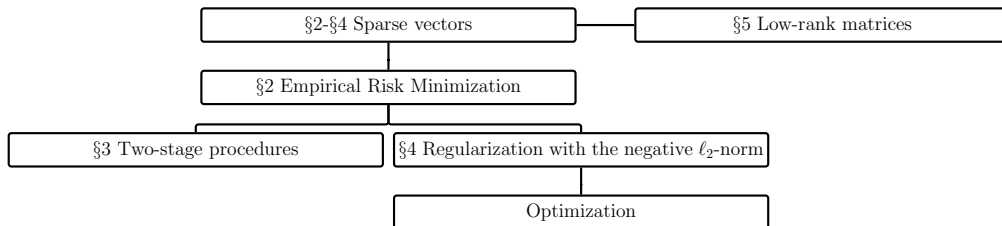
Relevant prior studies include the works of Larsson and Ugander (2011) and Shashanka, Raj and Smaragdis (2008), who discuss the aforementioned problem in the context of latent variable models for image and bag-of-words data. Larsson and Ugander (2011) propose a so-called pseudo-Dirichlet prior, akin to the log-penalty in Candes, Wakin and Boyd (2007). Shashanka, Raj and Smaragdis (2008) suggest using Shannon entropy as a regularizer. A conceptually different approach is pursued in Jojic, Saria and Koller (2011). Instead of the usual loss +  $\ell_1$ -norm formulation with the  $\ell_1$ -norm arising as the convex envelope of the  $\ell_0$ -norm on the unit  $\ell_\infty$ -ball, the authors work with the convex envelope of the loss +  $\ell_0$ -norm.

Finally, it is worth mentioning a line of research on sparse regression under linear inequality or equality constraints. Here, relevant works include those of James, Paulson and Rusmevichientong (2015), Lin et al. (2014), and Shi, Zhang and Li (2016). Meinshausen (2013) and Slawski and Hein (2013) study the case in which the constraint set is the non-negative orthant. It is shown that, under specific conditions, this constraint has similar effects to those of  $\ell_1$ -regularization.

With simplex constraints, this effect applies more broadly, as discussed in §2.

*Outline and contributions.* As a preliminary step, we provide a brief analysis of high-dimensional estimations under simplex constraints in §2. Such analyses provide valuable insights when designing sparsity-promoting schemes. Note that empirical risk minimization (ERM) and the elements of  $\Delta^p$  contained in a “high confidence set” for  $\beta^*$  (a construction inspired by the Dantzig selector of Candes and Tao (2007)) already enjoy nice statistical guarantees, including adaptation to sparsity under a restricted strong convexity condition that is weaker than that in Negahban et al. (2012). Next, we discuss strategies to improve on ERM, particularly with respect to the sparsity of the solution and support recovery. As a basic strategy, we consider simple two-stage procedures, thresholding and reweighted  $\ell_1$ -regularization on top of ERM (see §3).

As an alternative, we propose a novel regularization-based scheme in §4, in which  $\beta \mapsto 1/\|\beta\|_2^2$  serves as a relaxation of the  $\ell_0$ -norm on  $\Delta^p$ . This regularizer naturally extends to the case of positive semidefinite Hermitian matrices of unit trace, as discussed in §5. On the optimization side, the approach can be implemented using difference-of-convex (DC) programming (Pham Dinh and Le Thi (1997)). Unlike other forms of concave regularization, such as the SCAD, capped  $\ell_1$ , or MCP penalties (Zhang and Zhang (2013)) no parameter other than the regularization parameter needs to be specified. For this purpose, we employ a generic BIC-type criterion (Schwarz (1978); Kim, Kwon and Choi (2012)) with the aim of selecting the correct model (resp., rank, in the matrix case). The Supplementary Material (Li, Rangapuram and Slawski (2018)) contains all proofs, as well as numerical experiments on compressed sensing, density estimation, portfolio optimization and quantum state tomography that demonstrate the effectiveness of both the two-stage procedures and the regularization-based approach. The following orgchart provides a quick overview of the organization of the paper.



*Notation.* For the convenience of readers, we first present the essential notation. We denote  $\|\cdot\|_q$  for  $q \in [0, \infty]$ , as the usual  $\ell_q$ -norm or the Schatten  $\ell_q$ -norm, depending on the context, and  $\langle \cdot, \cdot \rangle$  as the usual Euclidean inner product. We use  $|\cdot|$  for the cardinality of a set. The support of  $v \in \mathbb{R}^d$  is denoted by  $S(v) = \{j : v_j \neq 0\}$ . For  $J \subseteq \{1, \dots, d\}$ , we let  $v_J = (v_j)_{j \in J}$ . We write  $\mathbf{I}(\cdot)$  for the indicator function. We denote  $\{e_1, \dots, e_d\}$  as the canonical basis of  $\mathbb{R}^d$ . For  $A \subseteq \mathbb{R}^d$ ,  $\Pi_A : \mathbb{R}^d \rightarrow A$  denotes the Euclidean projection on  $A$ . For the functions  $f(n)$  and  $g(n)$ , we write  $f(n) \gtrsim g(n)$  and  $f(n) \lesssim g(n)$  if  $f(n) \geq Cg(n)$  and  $f(n) \leq Cg(n)$ , respectively, for some constant  $C > 0$ . We write  $f(n) \asymp g(n)$  if both  $f(n) \gtrsim g(n)$  and  $f(n) \lesssim g(n)$ . We also use the Landau symbols  $O(\cdot)$  and  $o(\cdot)$ .

## 2. Simplex Constraint in High-dimensional Problems: Basic Analysis

Before designing schemes that promote sparsity under the constraint  $\beta \in \Delta^p$ , it is worth deriving basic performance bounds and establishing adaptivity to underlying sparsity when only simplex constraints are used for the estimation, without explicitly enforcing sparse solutions. Note that the constraint  $\beta \in \Delta^p$  is stronger than the  $\ell_1$ -ball constraint,  $\|\beta\|_1 \leq 1$ . As a result, ERM enjoys properties known from analyses of (unconstrained)  $\ell_1$ -regularized estimations, including the adaptivity to sparsity under certain conditions. This already sets a substantial limit on what can be achieved by sparsity-promoting schemes.

Let  $\{Z_i\}_{i=1}^n$  be independently and identically distributed (i.i.d.) copies of a random variable  $Z$  following a distribution  $P$  on a sample space  $\mathcal{Z} \subseteq \mathbb{R}^d$ . Let  $L : \mathbb{R}^p \times \mathcal{Z} \rightarrow \mathbb{R}$  be a loss function, such that  $\forall z \in \mathcal{Z}$ ,  $L(\cdot, z)$  is convex and differentiable. For  $\beta \in \mathbb{R}^p$ ,  $R(\beta) = \mathbf{E}[L(\beta, Z)]$  denotes the expected risk, and  $R_n(\beta) = n^{-1} \sum_{i=1}^n L(\beta, Z_i)$  denotes its empirical counterpart. The goal is to recover  $\beta^* = \operatorname{argmin}_{\beta \in \Delta^p} \mathbf{E}[L(\beta, Z)]$ . ERM yields  $\hat{\beta} \in \operatorname{argmin}_{\beta \in \Delta^p} R_n(\beta)$ . Figure 1 provides an overview of the key quantities and their relationships.

In addition to ERM, our analysis simultaneously covers all elements of the set

$$\mathcal{D}(\lambda) = \{\beta \in \Delta^p : \|\nabla R_n(\beta)\|_\infty \leq \lambda\}, \quad (2.1)$$

for suitably chosen  $\lambda \geq 0$ , as discussed below. The construction of  $\mathcal{D}(\lambda)$  is inspired by the constraint set of the Dantzig selector (Candes and Tao (2007)), which is extended to general loss functions by Lounici (2008); James and Radchenko (2009); Fan (2013). Elements of  $\mathcal{D}(\lambda)$  are shown to have performance comparable to  $\hat{\beta}$ . The set  $\mathcal{D}(\lambda)$  need not be convex, in general. For squared loss,

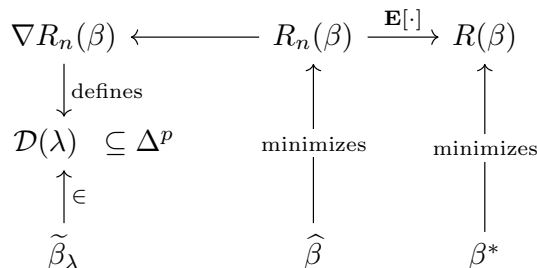


Figure 1. Diagram summarizing the relationships between the quantities employed in this study.

it becomes a convex polyhedron, which is nonempty as long as  $\lambda \geq \lambda_*$ , where  $\lambda_* = \|\nabla R_n(\beta^*)\|_\infty$ . In many settings of interest (cf., Lounici (2008); Negahban et al. (2012)), it can be shown that

$$\mathbf{P} \left( \lambda_* \gtrsim \sqrt{\frac{\log(p)}{n}} \right) = o(1) \quad \text{as } n \rightarrow \infty. \quad (2.2)$$

## 2.1. Excess risk

The first result bounds the excess risk of  $\hat{\beta}$  and  $\tilde{\beta}_\lambda$ , where in what follows,  $\tilde{\beta}_\lambda$  represents an arbitrary element of  $\mathcal{D}(\lambda)$  in (2.1).

**Proposition 1.** For  $\beta \in \mathbb{R}^p$ , let  $\psi_n(\beta) = R_n(\beta) - R(\beta)$  and  $\bar{\psi}_n(\beta) = \psi_n(\beta) - \psi_n(\beta^*)$ . For  $r > 0$ , let  $\mathbb{B}_1^p(r; \beta^*) = \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_1 \leq r\}$  denote the  $\ell_1$ -ball of radius  $r$  centered at  $\beta^*$  and  $\bar{\Psi}_n(r) = \sup\{|\bar{\psi}_n(\beta)| : \beta \in \mathbb{B}_1^p(r; \beta^*)\}$ . We then have

$$\begin{aligned}
R(\hat{\beta}) - R(\beta^*) &\leq \bar{\Psi}_n(\|\hat{\beta} - \beta^*\|_1) \leq \bar{\Psi}_n(2), \\
R(\tilde{\beta}_\lambda) - R(\beta^*) &\leq \bar{\Psi}_n(\|\tilde{\beta}_\lambda - \beta^*\|_1) + \lambda \|\tilde{\beta}_\lambda - \beta^*\|_1 \leq \bar{\Psi}_n(2) + 2\lambda.
\end{aligned}$$

The excess risk of ERM and points in  $\mathcal{D}(\lambda)$  can thus be bounded by controlling  $\bar{\Psi}_n(r)$ , the supremum of the empirical process  $\bar{\psi}_n(\beta)$  over all  $\beta$ , with  $\ell_1$ -distance at most  $r$  from  $\beta^*$ . This supremum is well studied in the literature on  $\ell_1$ -regularization. For example, for linear regression with a fixed or random sub-Gaussian design and sub-Gaussian errors, as well as for a Lipschitz loss (e.g., logistic loss), it can be shown that (van de Geer (2008))

$$\mathbf{P} \left( \bar{\Psi}_n(r) \gtrsim r \sqrt{\frac{\log(p)}{n}} \right) = o(1) \quad \text{as } n \rightarrow \infty. \quad (2.3)$$

Using  $\hat{\beta} \in \Delta^p$  and  $\tilde{\beta}_\lambda \in \Delta^p$ , choosing  $\lambda \asymp \lambda_*$ , and invoking (2.2), Proposition 1

yields that the excess risk of ERM and points in  $\mathcal{D}(\lambda)$  scale as  $O(\sqrt{\log(p)/n})$ . As a result, ERM and finding a point in  $\mathcal{D}(\lambda)$  constitute persistent procedures, in the sense of Greenshtein and Ritov (2004).

## 2.2. Adaptation to sparsity

Proposition 1 does not entail further assumptions on  $\beta^*$  or  $R_n$ . In this subsection, we suppose that  $\beta^* \in \Delta_0^p(s)$  and that  $R_n$  obeys a restricted strong convexity (RSC) condition, defined as follows. Consider the set

$$\mathcal{C}^\Delta(s) = \{\delta \in \mathbb{R}^p : \exists J \subseteq \{1, \dots, p\}, |J| \leq s \text{ s.t. } \mathbf{1}^\top \delta_{J^c} = -\mathbf{1}^\top \delta_J, \delta_{J^c} \succeq 0\}. \quad (2.4)$$

Observe that  $\{\beta - \beta^* : \beta \in \Delta^p\} \subseteq \mathcal{C}^\Delta(s)$ . For the next result, we require  $R_n$  to be strongly convex over  $\mathcal{C}^\Delta(s)$ .

**Condition 1.** *We say that the  $\Delta$ -RSC condition is satisfied for sparsity level  $1 \leq s \leq p$  and constant  $\kappa > 0$  if, for all  $\beta \in \Delta_0^p(s)$  and  $\delta \in \mathcal{C}^\Delta(s)$ ,*

$$R_n(\beta + \delta) - R_n(\beta) - \nabla R_n(\beta)^\top \delta \geq \kappa \|\delta\|_2^2.$$

Condition 1 is an adaptation of a condition employed in Negahban et al. (2012) for the analysis of (unconstrained)  $\ell_1$ -regularized ERM. Note that for squared loss, Condition 1 becomes the restricted eigenvalue condition in Bickel, Ritov and Tsybakov (2009), the range of validity of which has been investigated by, among others, Raskutti, Wainwright and Yu (2010); Rudelson and Zhou (2013); Lecue and Mendelson (2017). Our condition here is weaker, because the RSC condition in Negahban et al. (2012) is over the larger set

$$\mathcal{C}(\alpha, s) = \{\delta \in \mathbb{R}^p : \exists J \subseteq \{1, \dots, p\}, |J| \leq s \text{ s.t. } \|\delta_{J^c}\|_1 \leq \alpha \|\delta_J\|_1\},$$

for  $\alpha \geq 1$ . We can now state a second set of bounds.

**Proposition 2.** *Let the  $\Delta$ -RSC condition hold for sparsity level  $s$  and  $\kappa > 0$ . We then have*

$$\begin{aligned} \|\widehat{\beta} - \beta^*\|_2^2 &\leq \frac{4s\lambda_*^2}{\kappa^2}, & \|\widetilde{\beta}_\lambda - \beta^*\|_2^2 &\leq \frac{4s(\lambda + \lambda_*)^2}{\kappa^2}, \\ \|\widehat{\beta} - \beta^*\|_1 &\leq \frac{4s\lambda_*}{\kappa}, & \|\widetilde{\beta}_\lambda - \beta^*\|_1 &\leq \frac{4s(\lambda + \lambda_*)}{\kappa}. \end{aligned}$$

Invoking (2.2) and choosing  $\lambda \asymp \lambda_*$ , we recover the rates  $O(s \log(p)/n)$  for the squared  $\ell_2$ -error and  $O(s\sqrt{\log(p)/n})$  for the  $\ell_1$ -error, respectively. Combining the bounds on the  $\ell_1$ -error with (2.3) and Proposition 1, we obtain

$$R(\widehat{\beta}) - R(\beta^*) \lesssim \frac{s \log p}{n}, \quad R(\widetilde{\beta}_\lambda) - R(\beta^*) \lesssim \frac{s \log p}{n}.$$

The above rates are known to be minimax optimal for the parameter set  $\mathbb{B}_0^p(s)$  and squared loss (Ye and Zhang (2010)). Thus under the  $\Delta$ -RSC condition, there does not seem to be much room for improving over  $\widehat{\beta}$  and  $\widetilde{\beta}_\lambda$  as far as the  $\ell_1$ -error,  $\ell_2$ -error, and excess risk are concerned. An additional advantage of  $\widehat{\beta}$  is that it does not depend on any hyperparameters.

### 3. Two-stage Procedures

While  $\widehat{\beta}$  has appealing adaptation properties with regard to underlying sparsity,  $\|\widehat{\beta}\|_0$  may be significantly larger than the sparsity level  $s$ . Note that the  $\ell_2$ -bound of Proposition 2 yields that  $S(\widehat{\beta}) \supseteq S(\beta^*)$  as long as  $b_{\min}^* \gtrsim \lambda^* \sqrt{s}$ , where  $b_{\min}^* = \min\{\beta_j^* : j \in S(\beta^*)\}$ . If the aim is an estimator  $\widehat{\theta}$  that achieves support recovery, that is,  $S(\widehat{\theta}) = S(\beta^*)$ ,  $\widehat{\beta}$  needs to be further sparsified by a suitable form of post-processing. Here, we consider two schemes, namely thresholding and weighted  $\ell_1$ -regularization:

<b>Stage 1</b>	<b>Stage 2</b>	
Compute $\widehat{\beta}$	thresholding:	$\widehat{\beta}_\tau = (\widehat{\beta}_j \cdot \mathbf{I}(\widehat{\beta}_j \geq \tau))_{1 \leq j \leq p}$ (3.1)

	or weighted $\ell_1$ :	$\widehat{\beta}_\lambda^w \in \operatorname{argmin}_{\beta \in \Delta^p} R_n(\beta) + \lambda \langle w, \beta \rangle$ , (3.2)
--	------------------------	--

where  $\mathbf{I}(\cdot)$  denotes the indicator function and  $w = (w_j)_{j=1}^p$  are non-negative weights. We restrict ourselves to the common choice  $w_j = 1/\widehat{\beta}_j$  if  $\widehat{\beta}_j > 0$ , and  $w_j = +\infty$  otherwise (s.t.  $(\widehat{\beta}_\lambda^w)_j = 0$ ), for  $j = 1, \dots, p$ . Note that weighted  $\ell_1$ -regularization is often referred to as the “adaptive lasso” method (Zou (2006)).

While its primary purpose is model selection, thresholding (3.1) can optionally be complemented by a refitting step with fixed support, that is, ERM with the additional constraints  $\beta_j = 0 \ \forall j \notin S(\widehat{\beta}_\tau)$ .

A third approach is to ignore the unit sum constraint first, such that  $\ell_1$ -regularization has a sparsity-promoting effect, and then to divide the output by its sum as a simple way to satisfy the following constraint:

<b>Stage 1</b>	<b>Stage 2</b>	
$\widehat{\beta}_\lambda^{\ell_1} \in \operatorname{argmin}_{\beta \in \mathbb{R}_+^p} R_n(\beta) + \lambda \mathbf{1}^\top \beta$	Normalize:	$\widehat{\beta}_\lambda^{\ell_1} / (\mathbf{1}^\top \widehat{\beta}_\lambda^{\ell_1})$ . (3.3)

From the point of view of optimization, (3.3) offers several advantages. Non-negativity constraints alone tend to be easier to handle than simplex constraints. For projected gradient-type algorithms, the projection on the constraint set becomes trivial. Moreover, coordinate descent is applicable because non-negativity



constraints do not couple several variables (whereas simplex constraints do). Coordinate descent is one of the fastest algorithms for sparse estimation (Friedman, Hastie and Tibshirani (2010); Mazumder, Friedman and Hastie (2011)), particularly for large values of  $\lambda$ . On the other hand, from a statistical perspective, (3.3) is an ad hoc rather than a well-grounded approach. It is advisable to prefer  $\widehat{\beta}$  because it incorporates all given constraints into the optimization problem, which leads to a weaker RSC condition and eliminates the need to specify  $\lambda$  appropriately. Indeed, taking a large value of  $\lambda$  in (3.3) in order to obtain a highly sparse solution increases the bias and may lead to false negatives. In addition, (3.3) may also lead to false positives if the “irrepresentable condition” (Zhao and Yu (2006)) is violated. Our experimental results (cf., Supplementary Material, Li, Rangapuram and Slawski (2018)) confirm that (3.3) has a considerably larger estimation error than that of ERM.

*Model selection.* In this paragraph, we briefly discuss a data-driven approach for selecting the parameters  $\tau$  and  $\lambda$  in (3.1) and (3.2) when the aim is support recovery. It suffices to pick  $\tau$  from  $T = \{\widehat{\beta}_j\}_{j=1}^p$ , whereas for (3.2), we consider a finite set  $\Lambda \subset \mathbb{R}^+$ . We first obtain  $\{\widehat{\beta}_\tau, \tau \in T\}$  or  $\{\widehat{\beta}_\lambda^w, \lambda \in \Lambda\}$ , and then select one of the candidate models induced by the support set  $\{S(\widehat{\beta}_\tau), \tau \in T\}$  or  $\{S(\widehat{\beta}_\lambda^w), \lambda \in \Lambda\}$ , respectively. Model selection can be performed using a hold-out data set or an appropriate model selection criterion, such as the RIC in the case of squared loss (Foster and George (1994)). Specifically, let  $Z_i = (X_i, Y_i)$ , for  $i = 1, \dots, n$ , and suppose that

$$Y_i = X_i^\top \beta^* + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n. \tag{3.4}$$

Then, for  $S \subseteq \{1, \dots, p\}$ , the RIC is defined as

$$\text{RIC}(S) = \min_{\beta \in \mathbb{R}^p: \beta_{S^c} = 0} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^\top \beta)^2 + \frac{2\sigma^2 \log p}{n} |S|. \tag{3.5}$$

While computationally intractable in general, minimizing (3.5) has been shown to be model-selection consistent in high-dimensional regimes (Kim, Kwon and Choi (2012); Zhang and Zhang (2013)). Here, we minimize (3.5) over  $\{S(\widehat{\beta}_\tau)\}_{\tau \in T}$  or  $\{S(\widehat{\beta}_\lambda^w)\}_{\lambda \in \Lambda}$  only. The rationale is that support recovery is still achieved whenever the RIC is satisfied, provided that

$$S(\beta^*) \in \{S(\widehat{\beta}_\tau)\}_{\tau \in T} \tag{3.6}$$

$$\text{or } S(\beta^*) \in \{S(\widehat{\beta}_\lambda^w)\}_{\lambda \in \Lambda}. \tag{3.7}$$

Condition (3.6) is met if  $\min_{j \in S(\beta^*)} \widehat{\beta}_j > \max_{j \in S(\beta^*)^c} \widehat{\beta}_j$ , which can in turn be deduced from a bound on  $\|\widehat{\beta} - \beta^*\|_2$  (cf., Proposition 2) and a corresponding lower bound on  $b_{\min}^* = \min\{\beta_j^* : j \in S(\beta^*)\}$ . For weighted  $\ell_1$ -regularization, (3.7) is implied by a similar, albeit slightly more stringent condition.

**Proposition 3.** *Consider model (3.4) with  $\{X_i\}_{i=1}^n$  deterministic, such that  $(1/n) \sum_{i=1}^n X_{ij}^2 = 1$  for all  $j$ , and  $\widehat{\beta}_\lambda^w$  in (3.2) with  $R_n(\beta) = (1/2n) \sum_{i=1}^n (Y_i - X_i^\top \beta)^2$ . Then, (3.7) is satisfied with probability at least  $1 - O(p^{-1})$  if*

- i)  $\min_{j \in S(\beta^*)} \widehat{\beta}_j \gtrsim \max_{j \in S(\beta^*)^c} \widehat{\beta}_j$ ,
- ii)  $\Lambda \ni \lambda$  s.t.  $\lambda = \min_{j \in S(\beta^*)} \widehat{\beta}_j \lambda_0$  with  $\sigma \sqrt{\frac{\log(p)}{n}} \lesssim \lambda_0 \lesssim b_{\min}^*$ .

The constants hidden in  $\gtrsim$  and  $O(\cdot)$  are provided in the proof of the above statement.

On a practical note, we point out that consistent model selection based on the RIC (3.5) presumes knowledge of  $\sigma$ , or an estimator  $\widehat{\sigma}$  obeying at least  $\widehat{\sigma} \asymp \sigma$  (Kim, Kwon and Choi (2012)). We refer to Sun and Zhang (2012); Fan, Guo and Hao (2012); Dicker (2014); Reid, Tibshirani and Friedman (2016) for specific estimators  $\widehat{\sigma}$ .

#### 4. Regularization with the Negative $\ell_2$ -norm

A concern with ERM (optionally followed by a sparsification step) is that potential prior knowledge about sparsity is not incorporated into the estimation. The hope is that by taking sparsity into account, the guarantees of §2 can be improved. In particular, it may be possible to weaken Condition 1.

It turns out that any sparsity-promoting regularizer  $\Omega$  on  $\Delta^p$  cannot be convex. To see this, note that if  $\Omega$  is sparsity-promoting and homogeneous across coordinates, it should assign strictly smaller values to any of the vertices  $\{e_j\}_{j=1}^p$  of  $\Delta^p$  (which are maximally sparse) than to its barycentre (which is maximally dense); that is,

$$\Omega(e_j) < \Omega\left(\frac{\{e_1 + \dots + e_p\}}{p}\right), \quad j = 1, \dots, p. \quad (4.1)$$

However, (4.1) contradicts the convexity of  $\Omega$ , because by Jensen's inequality,

$$\Omega\left(\frac{\{e_1 + \dots + e_p\}}{p}\right) \leq \frac{\{\Omega(e_1) + \dots + \Omega(e_p)\}}{p}.$$

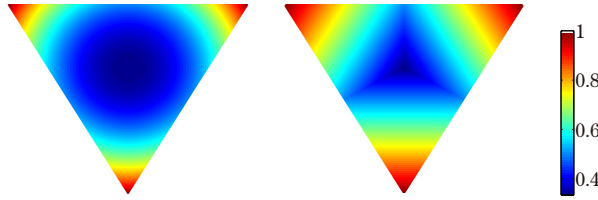


Figure 2. Contours of  $\beta \mapsto \|\beta\|_2^2$  (left) and  $\beta \mapsto \|\beta\|_\infty$  (right) on  $\Delta^3$ .

**4.1. Approach**

For  $0 \neq \beta \in \mathbb{R}^p$ , consider  $\Omega(\beta) = \|\beta\|_1^2 / \|\beta\|_2^2$ . Here,  $\Omega$  can be viewed as a “robust” measure of sparsity. We have  $\|\beta\|_0 \geq \Omega(\beta)$ , with equality holding iff  $\{|\beta_j|, j \in S(\beta)\}$  is constant. By “robustness” we mean that  $\Omega$  is small for vectors that have few entries of large magnitude, whereas the number of nonzero elements may be as large as  $p$ . From  $\|\beta\|_2^2 \leq \|\beta\|_\infty \|\beta\|_1$ , we have the alternative,  $\bar{\Omega}(\beta) = \|\beta\|_1 / \|\beta\|_\infty$ . As  $\|\beta\|_1 = 1 \forall \beta \in \Delta^p$ , we have

$$\frac{1}{\|\beta\|_\infty} \leq \frac{1}{\|\beta\|_2^2} \leq \|\beta\|_0 \quad \forall \beta \in \Delta^p. \tag{4.2}$$

The map  $\beta \mapsto 1/\|\beta\|_\infty$  is proposed as a sparsity-promoting regularizer on  $\Delta^p$  by Pilanci, Ghaoui and Chandrasekaran (2012). It yields a looser lower bound on  $\beta \mapsto \|\beta\|_0$  than that of the map  $\beta \mapsto 1/\|\beta\|_2^2$  advocated in the present work. Both lower bounds are sparsity-promoting on  $\Delta^p$  as indicated by Figure 2.

This lets us propose the following modifications of  $\hat{\beta}$  and  $\tilde{\beta}_\lambda$ , respectively,

$$\hat{\beta}_\lambda^{\ell_2} \in \operatorname{argmin}_{\beta \in \Delta^p} R_n(\beta) - \lambda \|\beta\|_2^2, \tag{4.3}$$

$$\tilde{\beta}_\lambda^{\ell_2} \in \operatorname{argmin}_{\beta \in \mathcal{D}(\lambda)} -\|\beta\|_2^2, \text{ with } \mathcal{D}(\lambda) \text{ as in (2.1)}. \tag{4.4}$$

Note the correspondence of (4.3) / (4.4) on the one hand, and the lasso (resp., Dantzig selector) on the other hand.

For (4.3), it appears to be better to use  $1/\|\beta\|_2^2$  rather than of  $-\|\beta\|_2^2$ , given (4.2). Eventually, this becomes a matter of parameterization. Although  $\beta \mapsto \|\beta\|_0$  is the canonical measure of sparsity,  $\beta \mapsto -1/\|\beta\|_0$  provides another measure. It is lower bounded by  $\beta \mapsto -1/\|\beta\|_2^2$ . We prefer the negative over the inverse, for computational reasons: the optimization problem in (4.3) is a DC program (Pham Dinh and Le Thi (1997)) and, hence, is more amenable to optimization. The problem in (4.4) is also a DC program if  $\mathcal{D}(\lambda)$  is convex. Note that for (4.4), minimizing the negative  $\ell_2$ -norm is equivalent to minimizing the inverse  $\ell_2$ -norm.

## 4.2. Least squares denoising

In order to show that the negative  $\ell_2$ -norm combined with simplex constraints promotes exactly sparse solutions, we elaborate on (4.3) in the simple setup of denoising. Let  $Z_i = \beta_i^* + \varepsilon_i$ , for  $i = 1, \dots, n = p$ , where  $\beta^* \in \Delta_0^n(s)$  and  $\{\varepsilon_i\}_{i=1}^n$  represents random noise. We consider squared loss, i.e.,  $L(\beta, Z_i) = (Z_i - \beta)^2$ ,  $i = 1, \dots, n$ . This yields the optimization problem

$$\min_{\beta \in \Delta^n} \frac{1}{n} \|\mathbf{Z} - \beta\|_2^2 - \lambda \|\beta\|_2^2, \quad \mathbf{Z} = (Z_i)_{i=1}^n. \quad (4.5)$$

As stated below, (4.5) can be recast as a Euclidean projection of  $\mathbf{Z}/\gamma$  on  $\Delta^n$ , where  $\gamma$  is a function of  $\lambda$ . Using this property, we derive conditions on  $\beta^*$  and  $\lambda$  such that  $\widehat{\beta}_\lambda^{\ell_2}$  achieves support recovery.

**Proposition 4.** *Consider (4.5) and suppose that  $z_{(1)} > \dots > z_{(n)}$ , where  $\{z_{(i)}\}_{i=1}^n$  denotes the ordered realizations of  $\{Z_i\}_{i=1}^n$ . For all  $\lambda \geq 1/n$ , we have  $\widehat{\beta}_\lambda^{\ell_2} = (\mathbf{I}(Z_i = z_{(1)}))_{i=1}^n$ . For all  $0 \leq \lambda < 1/n$ , we have  $\widehat{\beta}_\lambda^{\ell_2} = \operatorname{argmin}_{\beta \in \Delta^n} \|\mathbf{Z}/\gamma - \beta\|_2^2$ , where  $\gamma = 1 - n\lambda$ . Moreover, if  $2s \max_{1 \leq i \leq n} |\varepsilon_i|/n < \lambda < 1/n$  and  $b_{\min}^* > (n\lambda)/s + 2 \max_{1 \leq i \leq n} |\varepsilon_i|$ , we have  $S(\widehat{\beta}_\lambda^{\ell_2}) = S(\beta^*)$ .*

In particular, for  $\lambda = (1 + \delta)2s \max_{1 \leq i \leq n} |\varepsilon_i|/n$ , for any  $\delta > 0$ , the required lower bound on  $b_{\min}^*$  becomes  $4(1 + \delta) \max_{1 \leq i \leq n} |\varepsilon_i|$ . For the sake of reference, note that in the Gaussian sequence model with  $\varepsilon_i \sim N(0, \sigma^2/n)$  (cf., Johnstone (2013)), we have  $\max_{1 \leq i \leq n} |\varepsilon_i| \asymp \sqrt{\log(n)/n}$ .

The denoising problem (4.5) can be viewed as a least squares regression problem in which the design matrix is the identity matrix. For general design matrices, the analysis becomes more difficult, particularly because the optimization problem may be neither convex nor concave. In the latter case, the minimum is attained at one of the vertices of  $\Delta^p$ .

## 4.3. Optimization

Both (4.3) and (4.4) are nonconvex in general. Furthermore, maximizing the Euclidean norm over a convex set is NP-hard in general (Pardalos and Vavasis (1991)). To solve these two problems, we exploit the fact that both objectives are in DC form, that is, they can be represented as  $f(\beta) = g(\beta) - h(\beta)$ , with  $g$  and  $h$  both being convex. Linearizing  $-h$  at a given point yields a convex majorant of  $f$  that is tight at that point. Repeatedly minimizing the majorant yields an iterative procedure known as the concave-convex procedure (CCCP, Yuille and Rangarajan (2003)), which falls into the more general framework of

majorization-minimization (MM) algorithms (Lange, Hunter and Yang (2000)). When applied to (4.3) and (4.4), this approach yields Algorithm 1.

---

**Algorithm 1**


---

(4.3):  $\min_{\beta \in \Delta^p} R_n(\beta) - \lambda \|\beta\|_2^2$

**Initialization:**  $\beta^0 \in \Delta^p$

**repeat**  $\beta^{k+1} \in \operatorname{argmin}_{\beta \in \Delta^p} R_n(\beta) - 2 \langle \beta^k, \beta - \beta^k \rangle$

**until**  $R_n(\beta^{k+1}) - 2 \langle \beta^k, \beta^{k+1} \rangle = R_n(\beta^k)$

---

(4.4):  $\min_{\beta \in \mathcal{D}(\lambda)} -\|\beta\|_2^2$

**Initialization:**  $\beta^0 \in \mathcal{D}(\lambda)$

**repeat**  $\beta^{k+1} \in \operatorname{argmin}_{\beta \in \mathcal{D}(\lambda)} -2 \langle \beta^k, \beta - \beta^k \rangle$

**until**  $\langle \beta^k, \beta^{k+1} - \beta^k \rangle = 0$

---

For the second part of Algorithm 1 to be practical, we assume that  $\mathcal{D}(\lambda)$  is convex. The above algorithms can be shown to yield strict descent until convergence to a limit point satisfying the first-order optimality condition of problems (4.3)/(4.4). This is the content of the next proposition.

**Proposition 5.** *Let  $f$  denote the objective in (4.3) or (4.4). The elements of the sequence  $\{\beta^k\}_{k \geq 0}$  produced by Algorithm 1 satisfy  $f(\beta^{k+1}) < f(\beta^k)$  until convergence. Moreover, the limit satisfies the first-order optimality condition of the respective problem.*

An appealing feature of Algorithm 1 is that solving each subproblem in the repeat step involves only minor modifications to the computational approaches used for ERM (resp. finding a feasible point in  $\mathcal{D}(\lambda)$ ). With  $R_n$  assumed to be convex, ERM is a convex optimization problem. If  $R_n$  is also smooth, off-the-shelf algorithms such as interior point methods, projected gradient descent, and conditional gradient descent (Bertsekas (1999)) can be employed. For common nonsmooth losses, such as an absolute loss or hinge loss, ERM can be converted into a linear program. For the squared loss and absolute loss, specialized algorithms are proposed in Vila and Schniter (2014).

When selecting the parameter  $\lambda$  using a grid search, we suggest solving the associated instances of (4.3)/(4.4) from the smallest to the largest value of  $\lambda$ , using the solution from the current instance as the initial iterate for the next one. For the smallest value of  $\lambda$ , we recommend using  $\hat{\beta}$  and any point from  $\mathcal{D}(\lambda)$  as the initial iterate for (4.3) and (4.4), respectively. Running Algorithm 1 for formulation (4.4) has the advantage that all iterates are contained in  $\mathcal{D}(\lambda)$ , and thus enjoy at least the statistical guarantees of  $\tilde{\beta}_\lambda$  derived in §2. According to our

numerical results, formulation (4.3) achieves better performance (cf., supplement Li, Rangapuram and Slawski (2018)).

## 5. Extension to the Matrix Case

As pointed out in the introduction, there is a matrix counterpart to the aforementioned problem in which the object of interest is a low-rank Hermitian positive semidefinite matrix of unit trace. This set of matrices includes the density matrices of quantum systems (Nielsen and Chuang (2000)). The task of reconstructing such density matrices from so-called observables (e.g., noisy linear measurements) is termed quantum state tomography (Paris and Rehacek (2004)). In the past few years, quantum state tomography based on Pauli measurements has attracted considerable interest in the field of mathematical signal processing and statistics (Gross et al. (2010); Gross (2011); Koltchinskii (2011); Wang (2013); Cai et al. (2016)).

Specifically, the setup we employ throughout this section is as follows. Let  $\mathbb{H}^m = \{B \in \mathbb{C}^{m \times m} : B = B^H\}$  be the Hilbert space of complex Hermitian matrices with inner product  $\langle F, G \rangle = \text{tr}(FG)$ ,  $(F, G) \in \mathbb{H} \times \mathbb{H}$ , and, henceforth, let  $\|\cdot\|_q$ , for  $0 \leq q \leq \infty$ , denote the Schatten  $q$ -“norm” of a Hermitian matrix, defined as the  $\ell_q$ -norm of its eigenvalues. Here,  $\|\cdot\|_0$  denotes the number of nonzero eigenvalues, or equivalently, the rank. We suppose that the target  $B^*$  is contained in  $\Delta_0^m(r) := \mathbf{B}_0^m(r) \cap \Delta^m$ , where

$$\mathbf{B}_0^m(r) := \{B \in \mathbb{H}^m : \|B\|_0 \leq r\}, \quad \Delta^m := \{B \in \mathbb{H}^m : B \succeq 0, \text{tr}(B) = 1\}.$$

That is,  $B^*$  is also positive semidefinite, of unit trace, and has rank at most  $r$ . In low-rank matrix recovery, the Schatten 1-norm (typically referred to as the nuclear norm) is commonly used as a convex surrogate for the rank (Recht, Fazel and Parillo (2010)). Because the nuclear norm is constant over  $\Delta^m$ , a different strategy is needed to promote low-rankedness under that constraint. In the sequel, we carry over our treatment of the vector case to the matrix case. The analogies are mostly direct; at certain points, however, the matrix case yields additional complications, as detailed below. For simplicity, we restrict ourselves to the setup in which  $Z_i = (X_i, Y_i)$  are such that

$$Y_i = \langle X_i, B^* \rangle + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n, \quad (5.1)$$

with  $\{X_i\}_{i=1}^n \subset \mathbb{H}^m$ . Equivalently,

$$\mathbf{Y} = \mathcal{X}(B^*) + \varepsilon, \quad \mathbf{Y} = (Y_i)_{i=1}^n, \quad \varepsilon = (\varepsilon_i)_{i=1}^n,$$

where  $\mathcal{X} : \mathbb{H}^m \rightarrow \mathbb{R}^n$  is a linear operator defined by  $(\mathcal{X}(B))_i = \langle X_i, B \rangle$ ,  $B \in \mathbb{H}^m$ ,

for  $i = 1, \dots, n$ . We consider squared loss; that is, for  $B \in \Delta^m$ , the empirical risk is given by

$$R_n(B) = \frac{\|\mathbf{Y} - \mathcal{X}(B)\|_2^2}{n}.$$

*Basic estimators.* As basic estimators, we consider the empirical risk minimization given by  $\widehat{B} \in \operatorname{argmin}_{B \in \Delta^m} R_n(B)$ , as well as  $\widetilde{B}_\lambda$ , where  $\widetilde{B}_\lambda$  is any point in the set

$$\begin{aligned} \mathbf{D}(\lambda) &= \{B \in \Delta^m : \|\nabla R_n(B)\|_\infty \leq \lambda\} \\ &= \left\{ B \in \Delta^m : \frac{2}{n} \|\mathcal{X}^*(\mathcal{X}(B) - y)\|_\infty \leq \lambda \right\}, \end{aligned} \quad (5.2)$$

where  $\mathcal{X}^* : \mathbb{R}^n \rightarrow \mathbb{H}^m$  is the adjoint of  $\mathcal{X}$ . Both  $\widehat{B}$  and  $\widetilde{B}_\lambda$  adapt to the rank of  $B^*$  under a restricted strong convexity condition. For  $B \in \mathbf{B}_0^m(r)$ , let  $\mathbb{T}(B)$  be the tangent space of  $\mathbf{B}_0^m(r) \subset \mathbb{H}^m$  at  $B$  (see Definition 1 in Supplementary Material, Li, Rangapuram and Slawski (2018)), and let  $\Pi_V$  denote the projection on a subspace  $V$  of  $\mathbb{H}^m$ .

**Condition 2.** We say that the  $\Delta$ -RSC condition is satisfied for rank  $r$  and constant  $\kappa > 0$  if  $\forall \Phi \in \mathcal{K}^\Delta(r)$ , it holds that  $\|\mathcal{X}(\Phi)\|_2^2/n \geq \kappa \|\Phi\|_2^2$ , where

$$\begin{aligned} \mathcal{K}^\Delta(r) &= \{\Phi \in \mathbb{H}^m : \exists B \in \mathbf{B}_0^m(r) \text{ s.t.} \\ &\operatorname{tr}(\Pi_{\mathbb{T}(B)^\perp}(\Phi)) = -\operatorname{tr}(\Pi_{\mathbb{T}(B)}(\Phi)) \text{ and } \Pi_{\mathbb{T}(B)^\perp}(\Phi) \succeq 0\}. \end{aligned}$$

The  $\Delta$ -RSC condition is weaker than the corresponding condition employed in Negahban and Wainwright (2011), which, in turn, is weaker than the matrix RIP condition (Recht, Fazel and Parillo (2010)). The next statement parallels Proposition 2, asserting that the constraint  $B \in \Delta^m$  alone is strong enough to take advantage of low-rankedness.

**Proposition 6.** Suppose that the  $\Delta$ -RSC condition is satisfied for rank  $r$  and  $\kappa > 0$ . Set  $\lambda_* = 2\|\mathcal{X}^*(\varepsilon)\|_\infty/n$ , where  $\mathcal{X}^* : \mathbb{R}^n \rightarrow \mathbb{H}^m$  is the adjoint of  $\mathcal{X}$ . We then have

$$\begin{aligned} \|\widehat{B} - B^*\|_2^2 &\leq \frac{4s\lambda_*^2}{\kappa^2}, & \|\widetilde{B}_\lambda - B^*\|_2^2 &\leq \frac{4s(\lambda + \lambda_*)^2}{\kappa^2}, \\ \|\widehat{B} - B^*\|_1 &\leq \frac{4s\lambda_*}{\kappa}, & \|\widetilde{B}_\lambda - B^*\|_1 &\leq \frac{4s(\lambda + \lambda_*)}{\kappa}. \end{aligned}$$

*Obtaining solutions of low rank.* Although  $\widehat{B}$  may have a low estimation error, its rank can far exceed that of  $B^*$ , even though the extra nonzero eigenvalues of  $\widehat{B}$  tend to be small. The simplest approach to obtaining solutions of low rank is to

threshold the spectrum of  $\widehat{B} = \widehat{U}\widehat{\Phi}\widehat{U}^\top$  (the r.h.s. representing the usual spectral decomposition); that is,  $\widehat{B}_\tau = \widehat{U}\widehat{\Phi}_\tau\widehat{U}^\top$ , where  $\widehat{\Phi}_\tau = \text{diag}(\{\mathbf{I}(\widehat{\phi}_j \geq \tau)\widehat{\phi}_j\}_{j=1}^m)$  for a threshold  $\tau > 0$ . Similarly, we may use the following analog to weighted  $\ell_1$ -regularization:

$$\begin{aligned} \widehat{B}_w &= \widehat{U} \text{diag}(\{\widehat{\phi}_{w,j}\}_{j=1}^m) \widehat{U}^\top \\ \text{with } \widehat{\phi}_w &\in \underset{\phi \in \Delta^p}{\text{argmin}} \frac{1}{n} \|\mathbf{Y} - \mathcal{X}(\widehat{U} \text{diag}(\{\phi_j\}_{j=1}^m) \widehat{U}^\top)\|_2^2 + \lambda \langle w, \phi \rangle, \end{aligned} \quad (5.3)$$

for non-negative weights  $\{w_j\}_{j=1}^m$ , as in the vector case. Note that the matrix of eigenvectors  $\widehat{U}$  is kept fixed at the second stage; the optimization is only over the eigenvalues. Alternatively, we can consider optimization over  $\Delta^m$ , with regularizer  $B \mapsto \|B\|_w = \sum_{j=1}^m w_j \phi_j(B)$ , for eigenvalues  $\phi_1(B) \geq \dots \geq \phi_m(B) \geq 0$  of  $B$ , in decreasing order. However, from the point of view of optimization  $\|\cdot\|_w$  poses difficulties, including possible nonconvexity (depending on  $w$ ).

*Regularization with the negative  $\ell_2$ -norm.* An additional positive aspect about the regularization scheme proposed in §4 is that it allows a straightforward extension to the matrix case, including the algorithm used for optimization (Algorithm 1). In contrast, for regularization with the inverse  $\ell_\infty$ -norm, which can be reduced to  $p$  convex optimization problems in the vector case, no such reduction seems to be possible in the matrix case. The analogs of (4.3) and (4.4) are given by

$$\widehat{B}_\lambda^{\ell_2} \in \underset{B \in \Delta^m}{\text{argmin}} R_n(\beta) - \lambda \|B\|_2^2, \quad (5.4)$$

$$\widetilde{B}_\lambda^{\ell_2} \in \underset{B \in \mathcal{D}(\lambda)}{\text{argmin}} -\|B\|_2^2. \quad (5.5)$$

Algorithm 1 can be employed for optimization *mutatis mutandis*. In the vector case and for squared loss, formulations (4.3) and (4.4) are comparable in terms of their computational requirements: each minimization problem inside the repeat-loop becomes a quadratic (resp., a linear) program, with a comparable number of variables/constraints. In the matrix case, however, (5.4) appears to be preferable because the subproblems are better suited to the proximal gradient method. In contrast, the constraint set in (5.5) requires a more sophisticated approach.

*Denosing.* Negative  $\ell_2$ -regularization, together with the constraint set  $\Delta^m$  enforces a solution of low rank, as exemplified here in the special case of denosing



of a real-valued matrix (i.e.,  $B^* \in \mathbb{H}^m \cap \mathbb{R}^{m \times m}$ ) contaminated by Gaussian noise. Specifically, the sampling operator  $\mathcal{X}(\cdot) = (\langle X_i, \cdot \rangle)_{i=1}^n$ , for  $n = m(m+1)/2$ , is equal to the symmetric vectorization operator; that is

$$\begin{aligned} X_1 &= e_1 e_1^\top, \quad X_2 = \frac{e_1 e_2^\top + e_2 e_1^\top}{\sqrt{2}}, \dots, X_m = \frac{e_1 e_m^\top + e_m e_1^\top}{\sqrt{2}}, \quad X_{m+1} = e_2 e_2^\top, \dots, \\ X_{2m-1} &= \frac{e_2 e_m^\top + e_m e_2^\top}{\sqrt{2}}, \dots, X_{m(m+1)/2} = \frac{e_{m-1} e_m^\top + e_m e_{m-1}^\top}{\sqrt{2}}. \end{aligned} \quad (5.6)$$

The following proposition uses a result in random matrix theory of Peng (2012).

**Proposition 7.** *Let  $B^* \in \Delta_0^m(r) \cap \mathbb{R}^{m \times m}$  with eigenvalues  $\phi_1^* \geq \dots \geq \phi_r^* > 0$  and  $\phi_{r+1}^* = \dots = \phi_m^* = 0$ , let  $\mathcal{X}$  be defined according to (5.6), and let  $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_m/m)$ ,  $\mathbf{Y} = \mathcal{X}(B^*) + \varepsilon$ . Consider the optimization problem*

$$\min_{B \in \Delta^m} \frac{1}{n} \|\mathbf{Y} - \mathcal{X}(B)\|_2^2 - \lambda \|B\|_2^2,$$

with minimizer  $\widehat{B}_\lambda^{\ell_2}$ , and define  $\Upsilon = B^* + \mathcal{X}^*(\varepsilon)$ . Then, for all  $\lambda \geq 1/n$ , we have  $\widehat{B}_\lambda^{\ell_2} = u_1 u_1^\top$ , where  $u_1$  is the eigenvector of  $\Upsilon$  corresponding to its largest eigenvalue. For all  $0 \leq \lambda < 1/n$ , we have  $\widehat{B}_\lambda^{\ell_2} = \operatorname{argmin}_{B \in \Delta^m} \|\Upsilon/\gamma - B\|_2^2$ , where  $\gamma = 1 - n\lambda$ . Moreover, there exist constants  $c_0, c, C > 0$  so that if  $r < c_0 m$ ,  $\lambda \geq 6\sigma r/n$ , and  $\phi_r^* \geq 5\sigma + n\lambda/r$ , we have  $\|\widehat{B}_\lambda^{\ell_2}\|_0 = r$ , with probability at least  $1 - C \exp(-cm)$ .

In particular, for  $\lambda = (1 + \delta)6\sigma r/n$  for some  $\delta > 0$ , the required lower bound on  $\phi_r^*$  becomes  $11(1 + \delta)\sigma$ , which is proportional to the noise level of the problem, as follows from the proof of the proposition.

## 6. Conclusion

Simplex constraints are beneficial in high-dimensional estimations, empirically achieving lower estimation errors than when using  $\ell_1$ -norm regularization in place of the constraint. In order to enhance the sparsity of the solution, simple two-stage methods (i.e., thresholding and weighted  $\ell_1$ -regularization) are effective. A more principled way to incorporate sparsity is to use a suitable regularizer. We have pointed out that under simplex constraints, sparsity cannot be promoted by convex regularizers. We have therefore considered nonconvex alternatives, among which regularization using the negative  $\ell_2$ -norm turns out to be a natural approach, lending itself to a straightforward computational strategy. As an attractive feature, there is a direct and practical generalization to the matrix counterpart, in contrast to the two-stage methods.

## Supplementary Material

The Supplementary Material (Li, Rangapuram and Slawski (2018)) contains proofs of all statements, as well as extensive numerical results and simulations illustrating the central aspects of this work.

## Acknowledgments

The work was partially supported by NSF-DMS-1444124, NSF-III-1360971, and AFOSR-FA9550-13-1-0137. The work of Syama Sundar Rangapuram was also partially supported by the ERC starting grant NOLEPRO. The authors would like to thank Anastasios Kyrillidis for clarifications regarding step size selection for the iterative hard thresholding method discussed in the present work.

We thank the three reviewers for their careful reading and thoughtful comments, which led to several improvements to the paper.

## References

- Bertsekas, D. (1999). *Nonlinear Programming*. Athena Scientific.
- Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37**, 1705–1732.
- Blumensath, T. and Davies, M. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis* **27**, 265–274.
- Bunea, F., Tsybakov, A., Wegkamp, M. and Barbu, A. (2010). SPADES and mixture models. *The Annals of Statistics* **38**, 2525–2558.
- Cai, T., Kim, D., Wang, Y., Yuan, M. and Zhou, H. (2016). Optimal large-scale quantum state tomography with Pauli measurements. *The Annals of Statistics* **44**, 681–712.
- Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *The Annals of Statistics* **35**, 2313–2351.
- Candes, E., Wakin, M. and Boyd, S. (2007). Enhancing sparsity by reweighted  $\ell_1$ -minimization. *Journal of Fourier Analysis and Applications* **14**, 877–905.
- Dicker, L. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101**, 269–284.
- Fan, J. (2013). Features of big data and sparsest solution in high confidence set. Tech. rep., Department of Operations Research and Financial Engineering, Princeton University.
- Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultra-high dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 37–65.
- Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* **22**, 1947–1975.
- Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularized paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

- Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **6**, 971–988.
- Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory* **57**, 1548–1566.
- Gross, D., Liu, Y.-K., Flammia, S., Becker, S. and Eisert, J. (2010). Quantum State Tomography via Compressed Sensing. *Physical Review Letters* **105**, 150401–15404.
- James, G., Paulson, C. and Rusmevichientong, P. (2015). Penalized and Constrained Regression, manuscript, University of Southern California.
- James, G. and Radchenko, P. (2009). A generalized dantzig selector with shrinkage tuning. *Biometrika* **96**, 323–337.
- Johnstone, I. (2013). Gaussian estimation: Sequence and wavelet models. <http://statweb.stanford.edu/~imj/GE06-11-13.pdf>.
- Jojic, V., Saria, S. and Koller, D. (2011). Convex envelopes of complexity controlling penalties: the case against premature envelopment. In *International Conference on Artificial Intelligence and Statistics (AISTATS)* **15** of JMLR W&CP, 399–406.
- Keshava, N. (2003). A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal* **14**, 55–78.
- Kim, Y., Kwon, S. and Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research* **13**, 1037–1057.
- Koltchinskii, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics* **39**, 2936–2973.
- Kyriillidis, A., Becker, S., Cevher, V. and C.Koch (2013). Sparse projections onto the simplex. In *International Conference on Machine Learning (ICML)* **28** of JMLR W&CP, 235–243.
- Lange, K., Hunter, D. and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics* **9**, 1–20.
- Larsson, M. and Ugander, J. (2011). A concave regularization technique for sparse mixture models. In *Advances in Neural Information Processing Systems (NIPS)* **24**, 1890–1898.
- Lecue, G. and Mendelson, S. (2017). Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society* **19**, 881–904.
- Li, P., Rangapuram, S. and Slawski, M. (2018). Supplementary material of “methods for sparse and low-rank recovery under simplex constraint”.
- Lin, W., Shi, P., Feng, R. and Li, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101**, 785–797.
- Lounici, K. (2008). High-dimensional stochastic optimization with the generalized Dantzig estimator, arXiv:0811.2281.
- Markowitz, H. (1952). Portfolio selection. *Journal of Finance* **7**, 77–91.
- Mazumder, R., Friedman, J. and Hastie, T. (2011). *SparseNet*: Coordinate descent with non-convex penalties. *Journal of the American Statistical Association* **106**, 1125–1138.
- Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *The Electronic Journal of Statistics* **7**, 1607–1631.
- Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. (2012). A unified framework for high-dimensional analysis of  $M$ -estimators with decomposable regularizers. *Statistical Science* **27**, 538–557.
- Negahban, S. and Wainwright, M. (2011). Estimation of (near) low-rank matrices with noise

- and high-dimensional scaling. *The Annals of Statistics* **39**, 1069–1097.
- Nemirovski, A. (2000). *Ecole d’Ete de Probabilites de Saint-Flour XXVIII*. Springer, Ch. ‘Topics in non-parametric statistics’.
- Nielsen, M. and Chuang, I. (2000). *Quantum Computation and Quantum Information*. Cambridge University Press.
- Pardalos, P. and Vavasis, S. (1991). Quadratic programming with one negative eigenvalue is NP-hard. *Journal of Global Optimization* **1**, 15–22.
- Paris, M. and Rehacek, J. (Eds.) (2004). *Quantum State Estimation*. Springer.
- Peng, M. (2012). Eigenvalues of Deformed Random Matrices, arXiv:1205.0572.
- Pham Dinh, T. and Le Thi, H. (1997). Convex analysis approach to D. C. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica* **22**, 289–355.
- Pilanci, M., Ghaoui, L. E. and Chandrasekaran, V. (2012). Recovery of Sparse Probability Measures via Convex Programming. In *Advances in Neural Information Processing Systems (NIPS)* **25**, 2420–2428.
- Raskutti, G., Wainwright, M. and Yu, B. (2010). Restricted nullspace and eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research* **11**, 2241–2259.
- Recht, B., Fazel, M. and Parillo, P. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review* **52**, 471–501.
- Reid, S., Tibshirani, R. and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica* **26**, 35–67.
- Rudelson, M. and Zhou, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Transactions on Information Theory* **59**, 3434–3447.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.
- Shashanka, M., Raj, B. and Smaragdis, P. (2008). Sparse Overcomplete Latent Variable Decomposition of Counts Data. In *Advances in Neural Information Processing Systems (NIPS)* **20**, 1313–1320.
- Shen, J. and Li, P. (2018). A tight bound of hard thresholding. *Journal of Machine Learning Research* **18**, 1–42.
- Shi, P., Zhang, A. and Li, H. (2016). Regression analysis for microbiome compositional data. *The Annals of Applied Statistics* **10**, 1019–1040.
- Slawski, M. and Hein, M. (2013). Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization. *The Electronic Journal of Statistics* **7**, 3004–3056.
- Sun, T. and Zhang, C. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898.
- van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36**, 614–645.
- Vila, J. and Schniter, P. (2014). An empirical-Bayes approach to recovering linearly constrained non-negative sparse signals. *IEEE Transactions on Signal Processing* **62**, 4689–4703.
- Wang, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. *The Annals of Statistics* **41**, 2462–2504.
- Ye, F. and Zhang, C. (2010). Rate Minimality of the Lasso and Dantzig Selector for the  $\ell_q$  loss in  $\ell_r$  balls. *Journal of Machine Learning Research* **11**, 3519–3540.
- Yuille, A. and Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation* **15**, 915–936.

- Zhang, C. and Zhang, T. (2013). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27**, 576–593.
- Zhao, P. and Yu, B. (2006). On model selection consistency of the lasso. *Journal of Machine Learning Research* **7**, 2541–2567.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

10900 NE 8th St, Bellevue, WA 98004, USA.

E-mail: pingli98@gmail.com

Amazon Development Center, Krausenstrasse 38, 10117 Berlin, Germany.

E-mail: rangapur@amazon.de

George Mason University, 4400 University Drive, MS 4A7, Fairfax, VA 22030-4444, USA.

E-mail: mslawsk3@gmu.edu

(Received May 2016; accepted May 2018)

## Methods for Sparse and Low-Rank Recovery under Simplex Constraints

Ping Li, Syama Sundar Rangapuram, and Martin Slawski

*Baidu Research; Amazon Research; George Mason University*

### Supplementary Material

This supplementary file has two main constituents. Section S1 contains extensive set of numerical results; we follow a division into the vector and matrix case, respectively. The remaining sections contain proofs of Propositions 1 to 7.

## S1 Empirical results

We have conducted a series of simulations to compare the different methods considered herein and to provide additional support for several key aspects of the present work. Specifically, we study compressed sensing, least squares regression, mixture density estimation, and quantum state tomography based on Pauli measurements in the matrix case. The first two of these only differ by the presence respectively absence of noise. We also present a real data analysis example concerning portfolio optimization for

NASDAQ stocks based on weekly price data from 03/2003 to 04/2008.

### S1.1 Compressed sensing

We consider the problem of recovering  $\beta^* \in \Delta_0^p(s)$  from few random linear measurements  $Y_i = \langle X_i, \beta^* \rangle$ , where  $X_i$  has standard Gaussian entries,  $i = 1, \dots, n$ . In short,  $\mathbf{Y} = \mathbf{X}\beta^*$  with  $\mathbf{Y} = (Y_i)_{i=1}^n$  and  $\mathbf{X}$  having the  $\{X_i\}_{i=1}^n$  as its rows. Identifying  $\beta^*$  with a probability distribution  $\pi$  on  $\{1, \dots, p\}$ , we may think of the problem as recovering  $\pi$  from expectations  $Y_i = \sum_{j=1}^p (X_i)_j \pi(\{j\})$ . We here show the results for  $p = 500$ ,  $s = 50$  and  $n = cs \log(p/s)$  with  $c \in [0.8, 2]$  (cf. Figure 1). The target  $\beta^*$  is generated by selecting its support uniformly at random, drawing the non-zero entries randomly from  $[0, 1]$  and normalizing subsequently. This is replicated 50 times for each value of  $n$ .

Several approaches are compared for the given task, assuming squared loss  $R_n(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n$ :

'Feasible set': Note that ERM here amounts to finding a point in  $\mathcal{D}(0)$ . The output is used as initial iterate for 'L2', 'weighted L1', and 'IHT' below.

'L2':  $\ell_2$ -norm maximization (4.4) with  $\lambda = 0$ , i.e., over

$$\begin{aligned} \mathcal{D}(0) &= \{\beta \in \Delta^p : \mathbf{X}^\top(\mathbf{X}\beta - \mathbf{Y}) = 0\} \\ &= \{\beta \in \Delta^p : \mathbf{X}\beta = \mathbf{Y}\} \text{ with probability 1.} \end{aligned} \tag{S1.1}$$

'Pilanci': The method of Pilanci, Ghaoui, and Chandrasekaran (2012) that maximizes the  $\ell_\infty$ -norm over (S1.1).

'weighted L1': Weighted  $\ell_1$ -norm minimization (cf. §3) over (S1.1).

'IHT': Iterative hard threshold under simplex constraints (Kyriillidis et al., 2013). Regarding the step size used for gradient projection, we use the method in Kyriillidis and Cevher (2011) which empirically turned out to be superior compared to a constant step size. 'IHT' is run with the correct value of  $s$  and is hence given an advantage.

**Results.** Figure 1 visualizes the fractions of recovery out of 50 replications. A general observation is that the constraint  $\beta \in \Delta^p$  is powerful enough to reduce the required number of measurements considerably compared to  $2s \log(p/s)$  when using standard  $\ell_1$ -minimization without constraints. At this point, we refer to Donoho and Tanner (2005) who gave a precise asymptotic characterization of this phenomenon in the regime  $n/p \rightarrow c \in (0, 1)$  and  $s/n \rightarrow c' \in (0, 1)$ . When solving the feasibility problem, one does not explicitly exploit sparsity of the solution (even though the constraint implicitly does). Enforcing sparsity via 'Pilanci', 'IHT', 'L2'



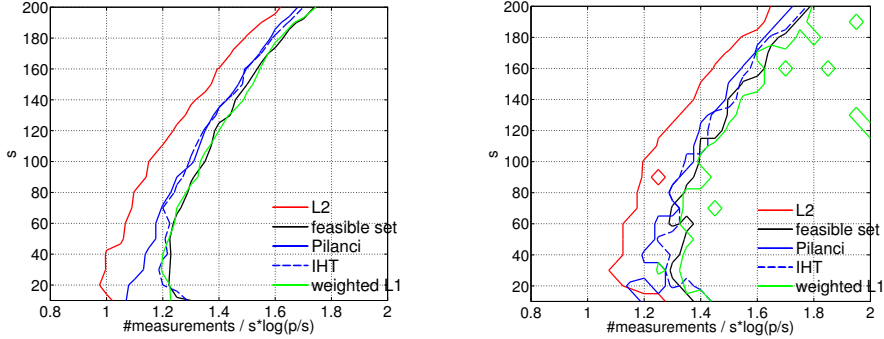


Figure 1: Contour plots of the empirical relative frequencies of exact recovery in dependency of the number of measurements (horizontal axis) and  $s$  (vertical axis). The left and right plot show the contour levels .75 and .99, respectively. Note that the smaller the area “left” to and “above” the curve, the better the performance.

further improves performance. The improvements achieved by ‘L2’ are most substantial and persist throughout all sparsity levels. ‘weighted L1’ does not consistently improve over the solution of the feasibility problem.

## S1.2 Least squares regression

We next consider the Gaussian linear regression model

$$Y_i = X_i^\top \beta^* + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n. \quad (\text{S1.2})$$

with the  $\{X_i\}_{i=1}^n$  as in the previous subsection. Put differently, the previous data-generating model is changed by an additive noise component. The target  $\beta^*$  is generated as before, with the change that the subvector  $\beta_{S(\beta^*)}^*$  corresponding to  $S(\beta^*)$  is projected on  $[b_{\min}^*, 1]^s \cap \Delta^s$  to ensure sufficiently

strong signal, where  $b_{\min}^* = \varrho\sigma\sqrt{2\log(p)/n}$  with  $\sigma = s^{-1}$  and  $\varrho = 1.7$  controlling the signal strength relative to the noise level  $\lambda_0 = \sigma\sqrt{2\log(p)/n}$ .

The following approaches are compared.

'ERM': Empirical risk minimization.

'Thres': 'ERM' followed by hard thresholding (cf. §3).

'L2-ERM': Regularized ERM with negative  $\ell_2$ -regularization (4.3). For the parameter  $\lambda$ , we consider a grid  $\Lambda$  of 100 logarithmically spaced points from 0.01 to  $\phi_{\max}(\mathbf{X}^\top \mathbf{X}/n)$ , the maximum eigenvalue of  $\mathbf{X}^\top \mathbf{X}/n$ . Note that for  $\lambda \geq \phi_{\max}(\mathbf{X}^\top \mathbf{X}/n)$ , the optimization problem (4.3) becomes concave and the minimizer must consequently be a vertex of  $\Delta^p$ , i.e., the solution is maximally sparse at this point, and it hence does not make sense to consider even larger values of  $\lambda$ . When computing the solutions  $\{\widehat{\beta}_\lambda^{\ell_2}, \lambda \in \Lambda\}$ , we use a homotopy-type scheme in which for each  $\lambda \in \Lambda$ , Algorithm 1 is initialized with the solution for the previous  $\lambda$ , using the output  $\widehat{\beta}$  of 'ERM' as initialization for the smallest value of  $\lambda$ .

'L2-D':  $\ell_2$ -norm maximization (4.4) over  $\mathcal{D}(C\lambda_0)$  with  $\lambda_0$  being the noise level defined above and  $C \in \{0.5, 0.55, \dots, 2\}$ . Algorithm 1 is initialized with  $\widehat{\beta}$  provided it is feasible. Otherwise, a feasible point is computed by

linear programming.

'weighted L1': The approach in (3.2). Regarding the regularization parameter, we follow van de Geer, Bühlmann, and Zhou (2013) who let  $\lambda = C\lambda_0^2$ .

We try 100 logarithmically spaced values between 0.1 and 10 for  $C$ .

'IHT': As above, again with the correct value of  $s$ . We perform a second sets of experiments though in which  $s$  is over-specified by different factors (1.2, 1.5, 2) in order to investigate the sensitivity of the method w.r.t. the choice of the sparsity level.

'L1': The approach (3.3), i.e., dropping the unit sum constraint and normalizing the output of the non-negative  $\ell_1$ -regularized estimator  $\widehat{\beta}_\lambda^{\ell_1}$ . We use  $\lambda = \lambda_0$  as recommended in the literature, cf. e.g. Negahban et al. (2012).

'oracle': ERM given knowledge of the support  $S(\beta^*)$ .

For 'Thres', 'L2-ERM' and other methods for which multiple values of a hyperparameter are considered, hyperparameter selection is done by minimizing the RIC as defined in §3 after evaluating each support set returned for a specific value of the hyperparameter.

**Results.** The results are summarized in Figures 2 and 3. Turning to the upper panel of Figure 2, the first observation is that 'L1' yields noticeably

S1. EMPIRICAL RESULTS

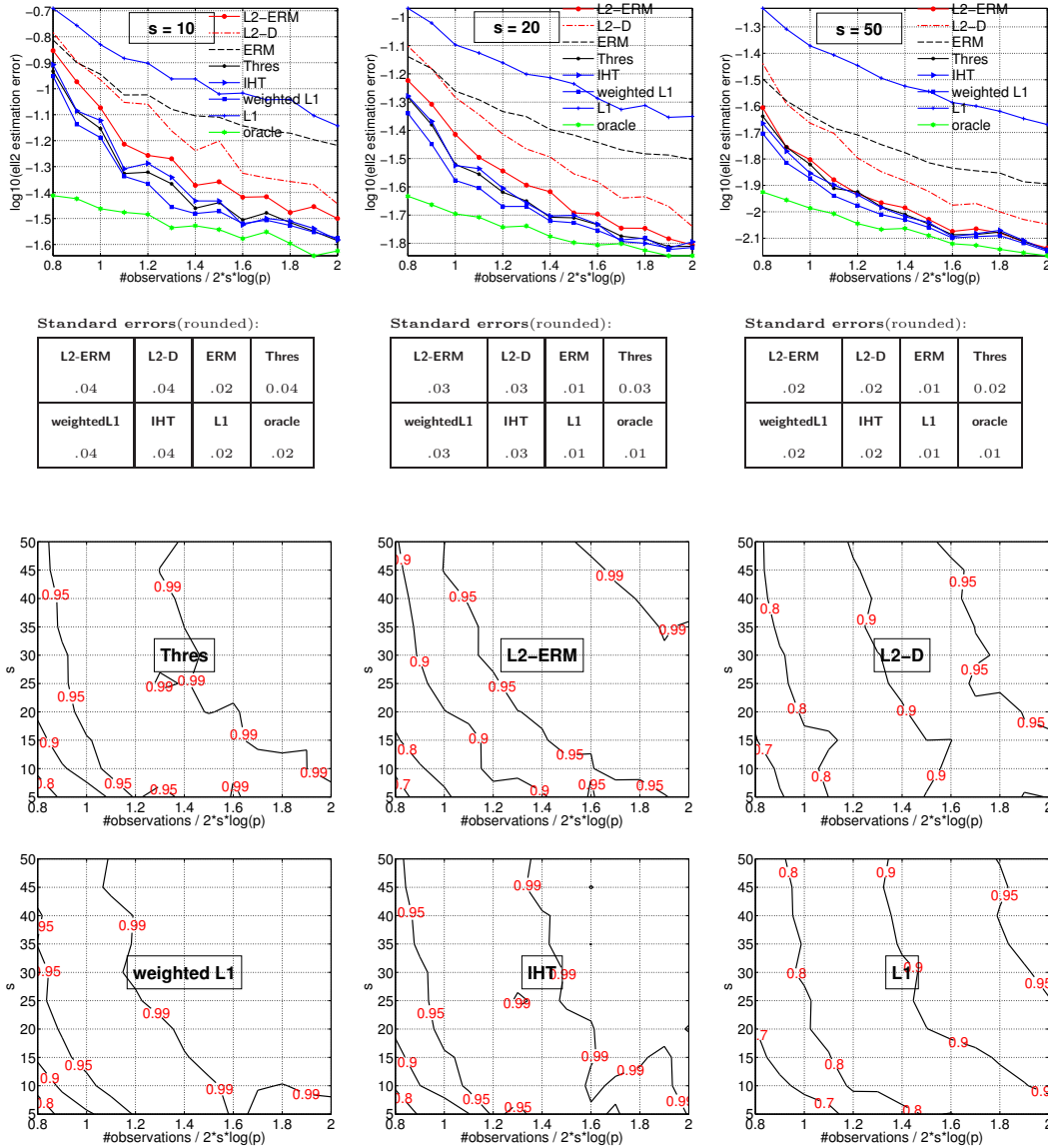


Figure 2: Upper panel: Average estimation errors  $\|\hat{\theta} - \beta^*\|_2$  (log<sub>10</sub> scale) in dependence of  $n$  over 50 trials for selected values of  $s$ . Here,  $\hat{\theta}$  is a placeholder for any of the estimators under consideration. Middle and Lower panel: contour plots of the average Matthew's correlation in dependence of  $n$  (horizontal axis) and  $s$  (vertical axis) for the contour levels 0.7, 0.8, 0.9, 0.95. Note that the smaller the area between the lower left corner of the plot and a contour line of a given level, the better the performance.

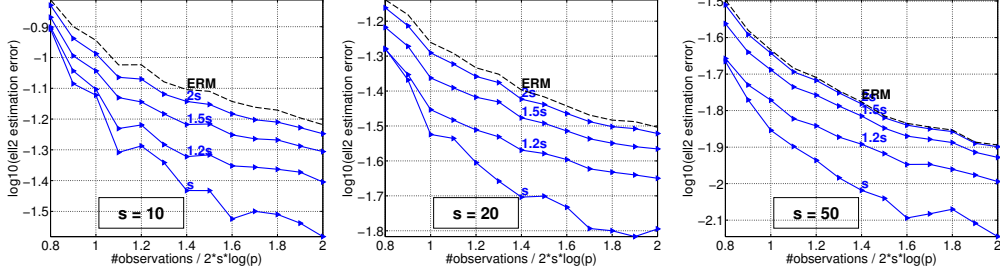


Figure 3: Sensitivity of 'IHT' w.r.t. the choice of  $s$ . The plots display error curves of IHT run with the correct value of  $s$  as appearing in Figure 2 as well with overspecification of  $s$  by the factors 1.2, 1.5, 2. The drop in performance is substantial: for  $2s$ , the improvement over ERM (here used as a reference) is only minor.

higher  $\ell_2$  estimation errors than 'ERM', which yields reductions roughly between a factor of  $10^{-1} \approx 0.79$  and  $10^{-2} \approx 0.63$ . A further reduction in error of about the same order is achieved by several of the above methods. Remarkably, the basic two-stage methods, thresholding and weighted  $\ell_1$ -regularization for the most part outperform the more sophisticated methods. Among the two methods based on negative  $\ell_2$ -regularization, 'L2-ERM' achieves better performance than 'L2-D'. We also investigate success in support recovery by comparing  $S(\hat{\theta})$  and  $S(\beta^*)$ , where  $\hat{\theta}$  represents any of the considered estimators, by means of Matthew's correlation coefficient (MCC) defined by

$$\text{MCC} = (\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}) / \{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})\}^{1/2},$$

with TP, FN etc. denoting true positives, false negatives etc. The larger the

criterion, which takes values in  $[0, 1]$ , the better the performance. The two lower panels of Figure 2 depict the MCCs in the form of contour plots, split by method. The results are consistent with those of the  $\ell_2$ -errors. The performance of 'weighted L1' and 'thres' improves respectively is on par with that of 'IHT' which is provided the sparsity level. Figure 3 reveals that this is a key advantage since the performance drops sharply as the sparsity level is over-specified by an increasing extent.

### S1.3 Density estimation

Let us recall the setup from the corresponding bullet in §1. For simplicity, we here suppose that the  $\{Z_i\}_{i=1}^n$  are i.i.d. random variables with density  $\phi_{\beta^*}$ , where for  $\beta \in \Delta^p$ ,  $\phi_\beta = \sum_{j=1}^p \phi_j \beta_j$  and  $\mathcal{F} = \{\phi_j\}_{j=1}^p$  is a given collection of densities. Specifically, we consider univariate Gaussian densities  $\phi_j = \phi_{\theta_j}$ , where  $\theta_j = (\mu_j, \sigma_j)$  contains mean and standard deviation,  $j = 1, \dots, p$ . As an example, one might consider  $p_0$  locations and  $K$  different standard deviations per location so that  $p = p_0 K$ , i.e.,  $\theta_{(k-1)p_0+l} = (\mu_l, \sigma_k)$ ,  $k = 1, \dots, K$ , and  $l = 1, \dots, p_0$ . This construction provides more flexibility compared to usual kernel density estimation where the locations equal the data points, a single bandwidth is used, and the coefficients  $\beta$  are all  $1/n$ . For large  $\mathcal{F}$ , sparsity in terms of the coefficients is common as a specific

target density can typically be well approximated by using an appropriate subset of  $\mathcal{F}$  of small cardinality.

As in Bunea et al. (2010), we work with the empirical risk

$$R_n(\beta) = \beta^\top Q\beta - 2c^\top \beta, \quad c = (\sum_{i=1}^n \phi_j(Z_i)/n)_{j=1}^p,$$

and  $Q = (\langle \phi_j, \phi_k \rangle)_{j,k=1}^p$ , where  $\langle f, g \rangle = \int_{\mathbb{R}} fg$  for  $f, g$  such that  $\|f\|, \|g\| < \infty$  with  $\|f\| = \langle f, f \rangle^{1/2}$ .

In our simulations, we let  $p_0 = 100$ ,  $K = 2$ ,  $\sigma_k = k$ ,  $k = 1, 2$ . The locations  $\{\mu_l\}_{l=1}^{p_0}$  are generated sequentially by selecting  $\mu_1$  randomly from  $[0, \delta]$ ,  $\mu_2$  from  $[\mu_1 + \delta, \mu_1 + 2\delta]$  etc. where  $\delta$  is chosen such that the 'correlations'  $\langle \phi_j, \phi_k \rangle / \|\phi_j\| \|\phi_k\| \leq 0.5$  for all  $(j, k)$  corresponding to different locations. An upper bound away from 1 is needed to ensure identifiability of  $S(\beta^*)$  from finite samples. Data generation, the methods compared, and the way they are run is almost identical to the previous subsections. Slight changes are made for  $S(\beta^*)$  (still selected uniformly at random, but it is ruled that any location is selected twice),  $b_{\min}^*$  ( $\varrho$  is set to 2) and hyperparameter selection. For the latter, a separate validation data set (also of size  $n$ ) is generated, and hyperparameters are selected as to minimize the empirical risk from the validation data.

**Results.** Figure 4 confirms once again that making use of simplex con-

straints yields markedly lower error than  $\ell_1$ -regularization followed by normalization (Bunea et al., 2010). 'L2-ERM' and 'weighted L1' perform best, improving over 'IHT' (which is run with knowledge of  $s$ ).

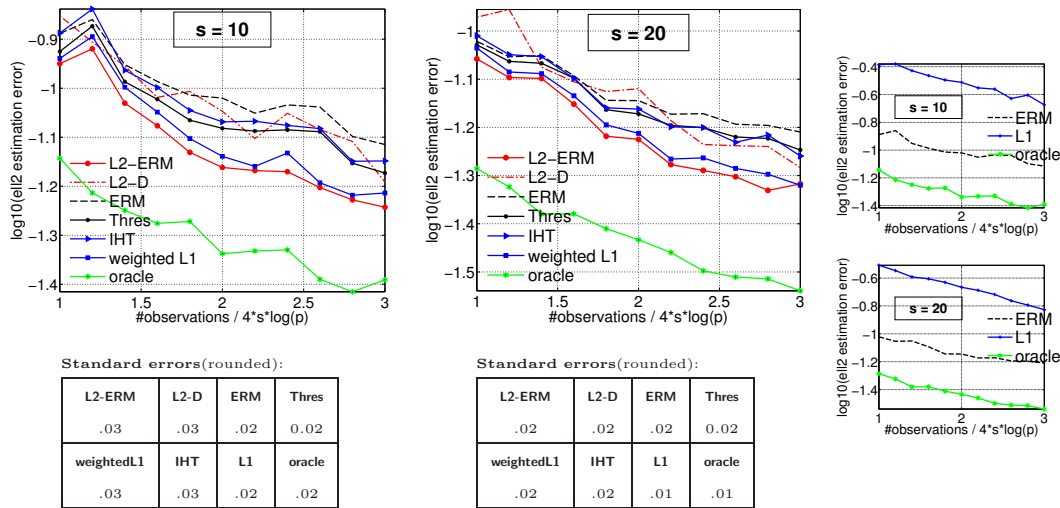


Figure 4: Average estimation errors  $\|\hat{\theta} - \beta^*\|_2$  for density estimation over 50 trials. Since the performance of 'L1' falls short of the rest of the competitors, whose differences we would like to focus on, 'L1' is compared to 'ERM' and 'oracle' in separate plots in the right column. Standard errors are smaller than 0.025 for all methods.

### S1.4 Portfolio Optimization

We use a data set available from [http://host.uniroma3.it/docenti/cesarone/datasetsw3\\_tardella.html](http://host.uniroma3.it/docenti/cesarone/datasetsw3_tardella.html) containing the weekly returns of  $p = 2196$  stocks in the NASDAQ index collected during 03/2003 and 04/2008 (264 weeks altogether). For each stock, the expected returns is



estimated as the mean return  $\hat{\mu}$  from the first four years (208 weeks). Likewise, the covariance of the returns is estimated as the sample covariance  $\hat{\Sigma}$  of the returns of the first four years. Given  $\hat{\mu}$  and  $\hat{\Sigma}$ , portfolio selection (without short positions) is based on the optimization problem

$$\min_{\beta \in \Delta^p} \beta^\top \hat{\Sigma} \beta - \tau \hat{\mu}^\top \beta \quad (\text{S1.3})$$

where  $\tau \in [0, \tau_{\max}]$  is a parameter controlling the trade-off between return and variance of the portfolio. Assuming that  $\hat{\mu}$  has a unique maximum entry,  $\tau_{\max}$  is defined as the smallest number such that the solution of (S1.3) has exactly one non-zero entry equal to one at the position of the maximum of  $\hat{\mu}$ . As observed in Brodie et al. (2009), the solution of (S1.3) tends to be sparse already because of the simplex constraint. Sparsity can be further enhanced with the help of the strategies discussed in this paper, treating (S1.3) as the empirical risk. We here consider 'L2-ERM', 'weighted L1', 'Thres' and 'IHT' for a grid of values for the regularization parameter ('L2-ERM' and 'weighted L1') respectively sparsity level ('L2-ERM' and 'Thres'). The solutions are evaluated by computing the Sharpe ratios (mean return divided by the standard deviation) of the selected portfolios on the return data of the last 56 weeks left out when computing  $\hat{\mu}$  and  $\hat{\Sigma}$ .

**Results.** Figure 5 displays the Sharpe ratios of the portfolios returned by these approaches in dependency of the  $\ell_2$ -norms of the solutions correspond-

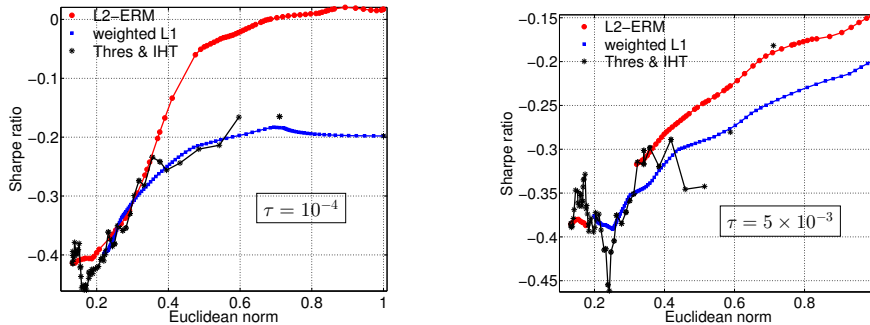


Figure 5: Sharpe ratios of the portfolios selected by 'L2-ERM', 'weighted L1', 'Thres' and 'IHT' on the hold-out portion of the NASDAQ data set in dependency of different choices for the regularization parameter/sparsity level (to allow for joint display, we use the  $\ell_2$ -norm as measure of sparsity on the horizontal axis). Left panel:  $\tau = 10^{-4}$ , Right panel:  $\tau = 5 \cdot 10^{-3}$ , cf. (S1.3). The results of 'Thres' and 'IHT' are essentially indistinguishable and are hence not plotted separately for better readability. Note that points that are too far away from each other with respect to the horizontal axis are not connected by lines.

ing to different choices of the regularization parameter respectively sparsity level and two values of  $\tau$  in (S1.3). One observes that promoting sparsity is beneficial in general. The regularization-based methods 'L2-ERM' and 'weighted L1' differ from 'IHT' and 'Thres' (whose results are essentially not distinguishable) in that the former two yield comparatively smooth curves. 'L2-ERM' achieves the best Sharpe ratios for a wide range of  $\ell_2$ -norms for both values of  $\tau$ .

### S1.5 Quantum State Tomography

We now turn to the matrix case of §5. The setup of this subsection is based on model (5.17), where the measurements  $\{X_i\}_{i=1}^n$  are chosen uniformly at random from the (orthogonal) Pauli basis of  $\mathbb{H}^m$  (here,  $m = 2^q$  for some integer  $q \geq 1$ ). For  $q = 1$ , the Pauli basis of  $\mathbb{H}^2$  is given by the following four matrices:

$$P_{1,1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P_{1,2} = \begin{pmatrix} 0 & -\sqrt{-1} \\ \sqrt{-1} & 0 \end{pmatrix}, \quad P_{1,3} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad P_{1,4} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

For  $q > 1$ , the Pauli basis  $\{P_{q,1}, \dots, P_{q,m^2}\}$  is constructed as the  $q$ -fold tensor product of  $\{P_{1,1}, P_{1,2}, P_{1,3}, P_{1,4}\}$ . The set of measurements is then given by  $\{P_{q,i}, i \in \mathcal{I}\}$ , where  $\mathcal{I} \subseteq \{1, \dots, m^2\}$ ,  $|\mathcal{I}| = n$ , is chosen uniformly at random. Pauli measurements are commonly used in quantum state tomography in order to recover the density matrix of a quantum state (see §5). In Gross et al. (2010), it is shown that if  $B^*$  is of low rank, it can be estimated accurately from few such random measurements by using nuclear norm regularization; the constraint  $B^* \in \Delta^m$  is not taken advantage of. Proposition 6 asserts that this constraint alone is well-suited for recovering matrices of low rank as long as the measurements satisfy a restricted strong convexity condition (Condition 2). It is shown in Liu (2011) that Pauli measurements satisfy the matrix RIP condition of Recht, Fazel, and Parillo (2010) as long as  $n \gtrsim mr \log^6(m)$ . Since the matrix RIP condition is stronger than Condition 2, Proposition 6 applies here. The requirement

on  $n$  is near-optimal: up to a polylogarithmic factor, it equals the “degrees of freedom” of the problem given by  $d = mr - r(r - 1)/2 \gtrsim mr$ , which is the dimension of the space  $\mathbb{T}(B^*) \subset \mathbb{H}^m$  (cf. Definition 1 in §S7 below).

### Noiseless measurements

In the first numerical study, we work with noiseless measurements. We fix  $m = 2^7$  and let  $r \in \{1, 2, 5, 10\}$  vary. The target is generated randomly as  $B^* = AA^\top$ , where  $A$  is an  $m \times r$  matrix, whose entries are drawn i.i.d. from  $N(0, 1)$ . The number of random Pauli measurements  $n$  are varied from  $2d$  to  $5d$  in steps of  $0.5d$ , where  $d$  equals the ‘degrees of freedom’ as defined above. For each possible combination of  $n$  and  $r$ , 50 trials are performed. The following three approaches for recovering  $B^*$  are compared.

‘Feasible set’: counterpart to ERM in the noiseless case: finding a point in

$$\mathbf{D}(0) = \{B \in \mathbf{\Delta}^m : \mathcal{X}^*(\mathcal{X}(B) - y) = 0\} = \{B \in \mathbf{\Delta}^m : \mathcal{X}(B) = y\}, \quad (\text{S1.4})$$

where the second identity follows from the fact that the Pauli matrices are unitary.

‘L2’: counterpart to (4.3)/(4.4) in the noiseless case, which amounts to maximizing the Schatten  $\ell_2$ -norm (i.e., Frobenius norm) over (S1.4). As initial iterate for Algorithm 1, the output from ‘feasible set’ is used.

'IHT': The matrix version of iterative hard thresholding under simplex constraints as proposed by Kyrillidis et al. (2013). Under the assumption that the rank of the target is known, one tries to solve directly the rank-constrained optimization problem  $\min_{B \in \Delta_0^m(r)} R_n(B)$  using projected gradient descent. Projections onto  $\Delta_0^m(r)$  can be efficiently computed using partial eigenvalue decompositions. We use a constant step size as in Kyrillidis et al. (2013). The output of 'feasible set' is used as initial iterate.

**Results.** Figure 6 shows a clear benefit of using  $\ell_2$ -norm maximization on top of solving the feasibility problem. For 'L2',  $2.5d$  measurements suffice to obtain highly accurate solutions, while 'feasible set' requires  $3.5d$  up to  $5d$  measurements. The performance of IHT falls in between the two other approaches even though the knowledge of  $r$  provides an extra advantage.

### Noisy measurements

We maintain the setup of the previous paragraph, but the measurements are now subject to additive Gaussian noise with standard deviation  $\sigma = 0.1$ . In order to adjust for the increased difficulty of the problem, the range for the number of measurements  $n$  is multiplied by the factor  $\log(m/r)$ . Our comparison covers the following methods.

'ERM': Empirical risk minimization, the counterpart to 'Feasible set' above.

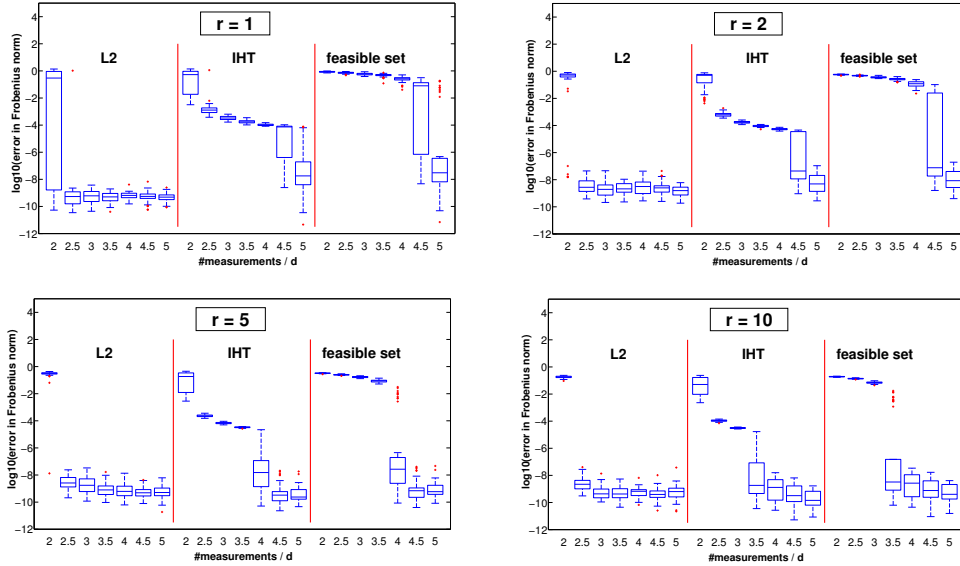


Figure 6: Boxplots of the errors  $\|\widehat{\Theta} - B^*\|_2$  (50 trials) in recovering  $B^*$  with respect to the Frobenius norm ( $\log_{10}$  scale) in dependence of the number of Pauli measurements ( $d = \text{'degrees of freedom'}$ ). Here,  $\widehat{\Theta}$  is representative for any of the three estimators under consideration.

'Thres': 'ERM' and eigenvalue thresholding, outlined below Proposition 6.

'L2-ERM': Regularized ERM with negative  $\ell_2$ -regularization (5.4). A grid search over 20 different values of the regularization parameter  $\lambda$  is performed analogously to the vector case.

'weighted L1': The approach in (5.3). The grid search for  $\lambda$  follows the vector case.

'IHT': As in the noiseless case.

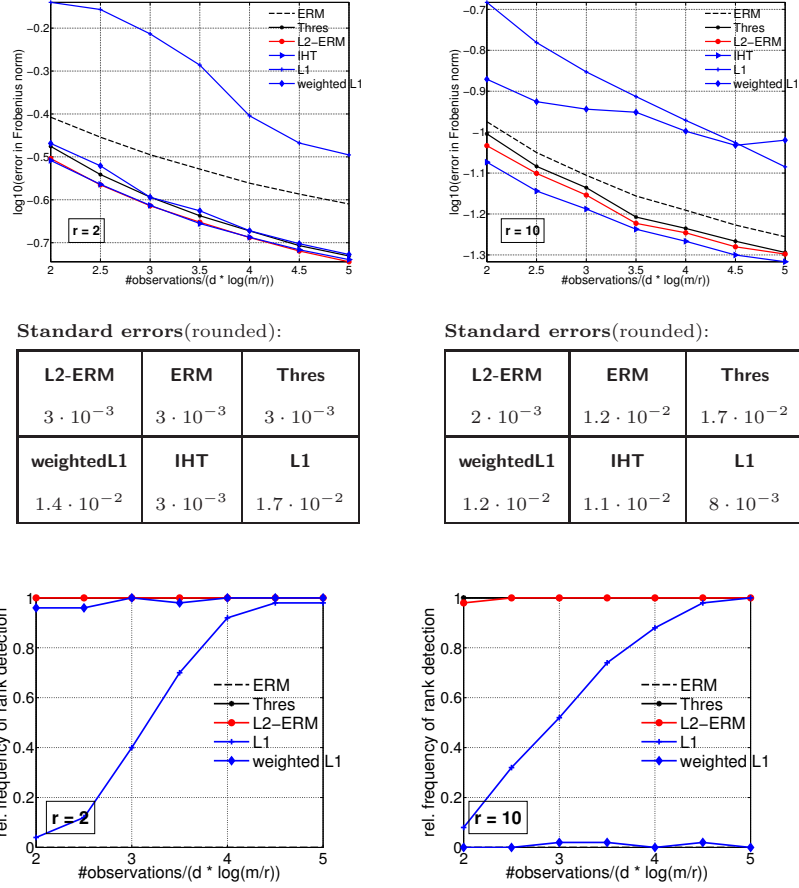


Figure 7: Bottom: Average estimation errors  $\|\hat{\Theta} - B^*\|_2$  over 50 trials ( $\log_{10}$ -scale) in dependence of the number of measurements ( $d =$  'degrees of freedom'). Top: Relative frequency of rank detection, i.e., of the event  $\{\|\hat{\Theta}\|_0 = \|B^*\|_0\}$ ; for 'IHT' this relative frequency is always one, which is not shown in the plots. Here,  $\hat{\Theta}$  is representative for any of the estimators under consideration.

'L1': In analogy to the counterpart (3.3) in the vector case, the unit trace constraint is dropped, and a nuclear-norm regularized empirical risk is minimized over the positive semidefinite cone. The result is then di-

vided by its trace. The regularization parameter is fixed to a single value  $\lambda_0 = 2\sigma\sqrt{\log(m)/n}$  according to the literature (Negahban and Wainwright, 2011; Koltchinskii, 2011).

For 'Thres', 'L2-ERM' and other methods for which multiple values of a hyperparameter are considered, hyperparameter selection is done by minimizing a RIC-type criterion. Specifically, for some estimate  $\hat{\Theta}_\lambda$  of  $B^*$ , we use

$$\text{sel}(\lambda) = R_n(\hat{\Theta}_\lambda) + \frac{C\sigma^2 \log(m^2) \|\hat{\Theta}_\lambda\|_0}{n}$$

The use of this criterion is justified in light of results in Klopp (2011) on trace regression with rank penalization. We have experimented with different choices of the constant  $C$ . Satisfactory results are achieved for  $C = 2^6$ , which is the choice underlying the results displayed in Figure 7. Once  $\lambda$  has been determined, the matrix of eigenvectors is fixed and the eigenvalues are re-fitted via least squares similar to (5.3).

**Results.** For the sake of brevity, we only display the results for  $r = 2, 10$  in Figures 7 and 8. 'IHT' achieves best performance even though the error curve of 'L2' is essentially identical for  $r = 2$ . Figure 8 indicates that 'IHT' is sensitive to the choice of  $r$ : over-specification by a factor of two



can lead to a performance that is significantly worse than 'Thres' and only slightly better than 'ERM'. Both 'L2' and 'Thres' are adaptive to the rank which is correctly recovered in almost all cases. In the matrix case, 'L2' improves over 'Thres' (as opposed to the vector case), possibly because for 'Thres' the eigenvectors remain unchanged compared to 'ERM', only the eigenvalues are modified. The performance of 'L1' clearly falls short of all other competitors, which underpins the importance of the unit trace constraint.

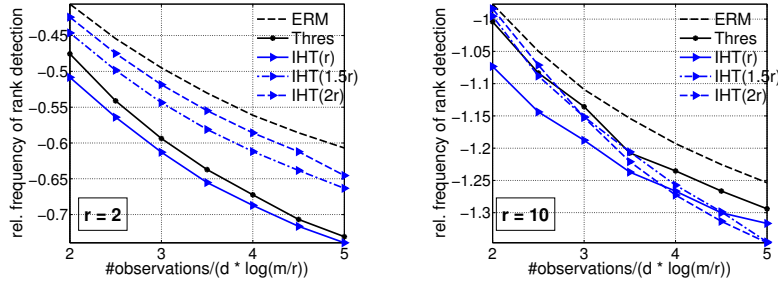


Figure 8: Sensitivity of 'IHT' w.r.t. the choice of  $r$ . The dashed-dotted and dashed lines show the average estimation errors when 'IHT' is run with  $1.5r$  and  $2r$ , respectively. The results of 'Thres' and 'ERM' serve as reference.

## S2 Proof of Proposition 1

By definition of  $\widehat{\beta}$ , we have

$$R_n(\widehat{\beta}) \leq R_n(\beta^*) \implies \{R_n(\widehat{\beta}) - R(\widehat{\beta})\} + R(\widehat{\beta}) \leq \{R_n(\beta^*) - R(\beta^*)\} + R(\beta^*).$$

The right hand side in turn implies that

$$\begin{aligned}
 R(\widehat{\beta}) &\leq R(\beta^*) + \sup_{\beta \in \mathbb{B}_1^p(\|\widehat{\beta} - \beta^*\|_1; \beta^*)} \underbrace{|\{R_n(\beta) - \{R_n(\beta^*)\} - \{R(\beta) - R(\beta^*)\}\}|}_{\overline{\psi}_n(\beta)} \\
 &= R(\beta^*) + \overline{\Psi}_n(\|\widehat{\beta} - \beta^*\|_1) \\
 &\leq R(\beta^*) + \overline{\Psi}_n(2),
 \end{aligned}$$

where the last inequality follows from  $\widehat{\beta} \in \Delta^p$ ,  $\beta^* \in \Delta^p$  and the triangle inequality.

We now turn to  $\widetilde{\beta}_\lambda$ . Consider the curve (segment)  $\gamma(t) = \beta^* + t(\widetilde{\beta}_\lambda - \beta^*)$  for  $t \in [0, 1]$  and the function  $g(t) = R_n(\beta^* + t(\widetilde{\beta}_\lambda - \beta^*))$ . Then  $g = R_n \circ \gamma$  is convex, as it is the composition of an affine and a convex function. Consequently, the derivative

$$g'(t) = \nabla R_n(\beta^* + t(\widetilde{\beta}_\lambda - \beta^*))^\top (\widetilde{\beta}_\lambda - \beta^*)$$

is non-decreasing. As a result, we have

$$\begin{aligned}
 R_n(\widetilde{\beta}_\lambda) - R_n(\beta^*) &= \int_0^1 \nabla R_n(\beta^* + t(\widetilde{\beta}_\lambda - \beta^*))^\top (\widetilde{\beta}_\lambda - \beta^*) dt \\
 &\leq \nabla R_n(\widetilde{\beta}_\lambda)^\top (\widetilde{\beta}_\lambda - \beta^*) \\
 &\leq \|\nabla R_n(\widetilde{\beta}_\lambda)\|_\infty \|\widetilde{\beta}_\lambda - \beta^*\|_1 \\
 &\leq \lambda \|\widetilde{\beta}_\lambda - \beta^*\|_1,
 \end{aligned}$$

where the first inequality follows from the definition and monotonicity property of  $g'$ , the second inequality is Hölder's inequality and the last

inequality follows from the definition of  $\tilde{\beta}_\lambda$ . Given the above bound on  $R_n(\tilde{\beta}_\lambda) - R_n(\beta^*)$ , the proof can be completed by following the scheme used for  $\hat{\beta}$ .

### S3 Proof of Proposition 2

Invoking the  $\Delta$ -RSC condition, we have

$$R_n(\hat{\beta}) - R_n(\beta^*) - \nabla R_n(\beta^*)^\top (\hat{\beta} - \beta^*) \geq \kappa \|\hat{\beta} - \beta^*\|_2^2,$$

On the other hand, by the definition of  $\hat{\beta}$

$$\begin{aligned} R_n(\hat{\beta}) - R_n(\beta^*) - \nabla R_n(\beta^*)^\top (\hat{\beta} - \beta^*) &\leq -\nabla R_n(\beta^*)^\top (\hat{\beta} - \beta^*) \\ &\leq \|\nabla R_n(\beta^*)\|_\infty \|\hat{\beta} - \beta^*\|_1. \end{aligned}$$

Combining these two bounds, we obtain that

$$\kappa \|\hat{\beta} - \beta^*\|_2^2 \leq \|\nabla R_n(\beta^*)\|_\infty \|\hat{\beta} - \beta^*\|_1.$$

This implies that

$$\begin{aligned} \|\hat{\beta} - \beta^*\|_2^2 &\leq \frac{\|\nabla R_n(\beta^*)\|_\infty^2}{\kappa^2} \left( \frac{\|\hat{\beta} - \beta^*\|_1}{\|\hat{\beta} - \beta^*\|_2} \right)^2 \leq \frac{4s\lambda_*^2}{\kappa^2}, \\ \|\hat{\beta} - \beta^*\|_1 &\leq \frac{\|\nabla R_n(\beta^*)\|_\infty}{\kappa} \left( \frac{\|\hat{\beta} - \beta^*\|_1}{\|\hat{\beta} - \beta^*\|_2} \right)^2 \leq \frac{4s\lambda_*}{\kappa}, \end{aligned}$$

where  $\lambda_* = \|\nabla R_n(\beta^*)\|_\infty$ . The rightmost inequalities follow from the fact that  $\widehat{\beta} - \beta^* \in \mathcal{C}^\Delta(s)$  and hence  $\|\widehat{\beta}_{S(\beta^*)^c}\|_1 \leq \|\widehat{\beta}_{S(\beta^*)} - \beta_{S(\beta^*)}^*\|_1$  so that

$$\begin{aligned} \|\widehat{\beta} - \beta^*\|_1 &= \|\widehat{\beta}_{S(\beta^*)} - \beta_{S(\beta^*)}^*\|_1 + \|\widehat{\beta}_{S(\beta^*)^c}\|_1 \\ &\leq 2\|\widehat{\beta}_{S(\beta^*)} - \beta_{S(\beta^*)}^*\|_1 \leq 2\sqrt{s}\|\widehat{\beta}_{S(\beta^*)} - \beta_{S(\beta^*)}^*\|_2. \end{aligned}$$

We now turn to  $\widetilde{\beta}_\lambda$ . Starting from

$$R_n(\widetilde{\beta}_\lambda) - R_n(\beta^*) - \nabla R_n(\beta^*)^\top (\widetilde{\beta}_\lambda - \beta^*) \geq \kappa \|\widetilde{\beta}_\lambda - \beta^*\|_2^2,$$

and using the upper bound on  $R_n(\widetilde{\beta}_\lambda) - R_n(\beta^*)$  as derived in the proof of Proposition 1, we obtain

$$\begin{aligned} \kappa \|\widetilde{\beta}_\lambda - \beta^*\|_2^2 &\leq \|\nabla R_n(\widetilde{\beta}_\lambda)\|_\infty \|\widetilde{\beta}_\lambda - \beta^*\|_1 + \|\nabla R_n(\beta^*)\|_\infty \|\widetilde{\beta}_\lambda - \beta^*\|_1 \\ &\leq (\lambda + \lambda_*) \|\widetilde{\beta}_\lambda - \beta^*\|_1, \end{aligned}$$

Arguing similarly as for  $\widehat{\beta}$ , it follows that

$$\|\widetilde{\beta}_\lambda - \beta^*\|_2^2 \leq \frac{4s(\lambda + \lambda_*)^2}{\kappa^2}, \quad \|\widetilde{\beta}_\lambda - \beta^*\|_1 \leq \frac{4s(\lambda + \lambda_*)}{\kappa}.$$

## S4 Proof of Proposition 3

We fix notation first. We let  $S = S(\beta^*)$ ,  $\mathbf{Y} = (Y_i)_{i=1}^n$  and  $\varepsilon = (\varepsilon_i)_{i=1}^n$ .

The matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$  has the  $\{X_i\}_{i=1}^n$  as its rows, and  $\mathbf{X}_S$ ,  $\mathbf{X}_{S^c}$  denote the column submatrices corresponding to  $S$  respectively  $S^c$ . Accordingly, we let

$\Sigma_{SS} = \frac{1}{n} \mathbf{X}_S^\top \mathbf{X}_S$ ,  $\Sigma_{S^c S^c} = \frac{1}{n} \mathbf{X}_{S^c}^\top \mathbf{X}_{S^c}$  and  $\Sigma_{S^c S} = \frac{1}{n} \mathbf{X}_{S^c}^\top \mathbf{X}_S$ . We recall that

$w = (w_j)_{j=1}^p$  with  $w_j = 1/\widehat{\beta}_j$ ,  $j = 1, \dots, p$ , so that  $\|w_S\|_\infty = 1/\min_{j \in S} \widehat{\beta}_j$ .

Moreover, we define

$$\phi_S = \min_{\|v\|_2=1} v^\top \Sigma_{SS} v, \quad \iota_S = \|\Sigma_{S^c S} \Sigma_{SS}^{-1}\|_\infty, \quad \varrho_{S,w} = \frac{\mathbf{1}^\top \Sigma_{SS}^{-1} \frac{w_S}{\|w_S\|_\infty}}{\mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{1}}, \quad (\text{S4.1})$$

where for a matrix  $M$ ,  $\|M\|_\infty = \max_{\|v\|_\infty \leq 1} \|Mv\|_\infty$ . Consider the optimization problems

$$\begin{aligned} (\circ) \quad & \min_{\beta \in \Delta^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / (2n) + \lambda \langle w, \beta \rangle, \\ (\bullet) \quad & \min_{\beta: \mathbf{1}^\top \beta_S = 1, \beta_{S^c} = 0} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 / (2n) + \lambda \langle w, \beta \rangle. \end{aligned} \quad (\text{S4.2})$$

Let  $\bar{\beta}$  denote the minimizer of  $(\bullet)$ . In the sequel, it will be verified that under the stated conditions  $\bar{\beta}$  also minimizes  $(\circ)$ . It follows from the KKT conditions of  $(\circ)$  that it suffices to show that

$$\begin{aligned} \text{I)} \quad & \bar{\beta}_S \succ 0, \\ \text{II)} \quad & \frac{1}{n} \mathbf{X}_{S^c}^\top (\mathbf{X}_S \bar{\beta}_S - \mathbf{Y}) \succ \bar{\mu} \mathbf{1} - \lambda w_{S^c}, \quad \bar{\mu} := \underbrace{-\frac{\mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{X}_S^\top \varepsilon / n}{\mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{1}}}_{=:\bar{\mu}_0} + \lambda \|w_S\|_\infty \varrho_{S,w}, \end{aligned} \quad (\text{S4.3})$$

where  $\succ, \succeq$  etc. denote component-wise inequality and  $\bar{\mu}$  is the optimal value of the Lagrangian multiplier associated with the constraint  $\mathbf{1}^\top \beta_S = 1$  in  $(\bullet)$ . Direct calculations show that I) holds if

$$\begin{aligned} b_{\min}^* &> T_\varepsilon + \lambda \|w_S\|_\infty \|\Sigma_{SS}^{-1}\|_\infty (1 + \varrho_{S,w}), \\ T_\varepsilon &:= \|\Sigma_{SS}^{-1} \mathbf{X}_S^\top \varepsilon / n\|_\infty + \|\bar{\mu}_0 \Sigma_{SS}^{-1} \mathbf{1}\|_\infty. \end{aligned} \quad (\text{S4.4})$$

Let  $\mathcal{P}$  denote the projection onto the column space of  $\mathbf{X}_S$ . Re-arranging II) in (S4.3) then yields

$$\lambda w_{S^c} \succ \lambda \Sigma_{S^c S} \Sigma_{SS}^{-1} w_S + \lambda \|w_S\|_\infty \varrho_{S,w} (\mathbf{1} - \Sigma_{S^c S} \Sigma_{SS}^{-1} \mathbf{1}) + \bar{\mu}_0 (\mathbf{1} - \Sigma_{S^c S} \Sigma_{SS}^{-1} \mathbf{1}) + \mathbf{X}_{S^c}^\top (I - \mathcal{P}) \varepsilon / n.$$

By upper bounding the right hand side component-wise, we obtain that II) in (S4.3) is implied by

$$\lambda \min_{j \in S^c} w_j > 2\lambda \max(\varrho_{S,w}, 1) (1 + \iota_S) \|w_S\|_\infty + T'_\varepsilon, \quad (\text{S4.5})$$

$$T'_\varepsilon := \|\bar{\mu}_0 (\mathbf{1} - \Sigma_{S^c S} \Sigma_{SS}^{-1} \mathbf{1}) + \mathbf{X}_{S^c}^\top (I - \mathcal{P}) \varepsilon / n\|_\infty,$$

with  $\iota_S$  as in (S4.1). Consider now the event

$$\mathcal{E} = \{T'_\varepsilon \leq \lambda \max(\varrho_{S,w}, 1) (1 + \iota_S) \|w_S\|_\infty\}.$$

Note that

$$\mathcal{E} \supseteq \{T''_\varepsilon \leq \lambda \|w_S\|_\infty\}, \quad T''_\varepsilon := |\bar{\mu}_0| + \|\mathbf{X}_{S^c}^\top (I - \mathcal{P}) \varepsilon / n\|_\infty. \quad (\text{S4.6})$$

Inserting  $\lambda = \lambda_0 \|w_S\|_\infty$  into (S4.4) and (S4.6) with  $\lambda_0$  still to be determined, we obtain the events

$$\{b_{\min}^* > T_\varepsilon + \lambda_0 \|\Sigma_{SS}^{-1}\|_\infty (1 + \varrho_{S,w})\}, \quad \{T''_\varepsilon \leq \lambda_0\}. \quad (\text{S4.7})$$

(A) Regarding  $T''_\varepsilon$ , observe that from the definition of  $\bar{\mu}_0$  in (S4.3), we get  $\bar{\mu}_0 \sim N(0, \frac{\sigma^2}{n} \{\mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{1}\}^{-1})$ . Indeed, using that  $\frac{1}{n} \sum_{i=1}^n X_{ij}^2 = 1 \forall j$  by assumption, which implies that  $\text{tr}(\Sigma_{SS}) = s$ , one shows that  $\{\mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{1}\}^{-1} \leq \frac{1}{s} \max_{\|v\|_2 \leq 1} v^\top \Sigma_{SS} v \leq \frac{1}{s} \text{tr}(\Sigma_{SS}) = 1$ . Likewise, we note that each component of  $\mathbf{X}_{S^c}^\top (I - \mathcal{P}) \varepsilon / n$  is a Gaussian random variable with variance at

most  $\sigma^2/n$ . Applying a standard maximal inequality for finite collections of Gaussian random variables (cf., e.g., Appendix A in Slawski and Hein (2013)), choosing  $\lambda_0 \geq (1 + \eta)\sqrt{2 \log(p)/n}$  for  $\eta \geq 0$  yields that the event  $\{T_\varepsilon'' \leq \lambda_0\}$  in (S4.7) holds with probability at least  $1 - 2p^{-\eta^2}$ .

(B) We now turn to the first event in (S4.7) which entails closer examination of  $T_\varepsilon$  in (S4.4). First, each component of  $\Sigma_{SS}^{-1} \mathbf{X}_S^\top \varepsilon/n$  is a Gaussian random variable with variance at most  $\phi_S^{-1} \sigma^2/n$ , where  $\phi_S$  is given in (S4.1). Second, using that  $\|\Sigma_{SS}^{-1} \mathbf{1}\|_\infty \leq \|\Sigma_{SS}^{-1} \mathbf{1}\|_2 = (\mathbf{1}^\top \Sigma_{SS}^{-2} \mathbf{1})^{1/2}$  and further that  $(\mathbf{1}^\top \Sigma_{SS}^{-2} \mathbf{1} / \mathbf{1}^\top \Sigma_{SS}^{-1} \mathbf{1})^{1/2} \leq \phi_S^{-1/2}$ , we obtain that the second term in  $T_\varepsilon$  is distributed as the absolute value of a Gaussian random variable with variance at most  $\phi_S^{-1} \sigma^2/n$ . Invoking the maximal inequality as used in the above paragraph (A), we conclude that the event  $\{T_\varepsilon \leq \phi_S^{-1/2} \lambda_0\}$  holds with probability at least  $1 - 2p^{-\eta^2}$ . Combining this with (S4.7) and (S4.4), we obtain that the event  $\{\bar{\beta}_S \succ 0\}$  in (S4.3) holds with probability at least  $1 - 4p^{-\eta^2}$  if  $b_{\min}^* \geq \lambda_0(\phi_S^{-1/2} + \|\Sigma_{SS}^{-1}\|_\infty(1 + \varrho_{S,w}))$ .

(C) Lastly, we inspect the condition in (S4.5) conditional on the event  $\mathcal{E}$  specified below (S4.5). Substituting  $\|w_S\|_\infty = 1/\min_{j \in S} \hat{\beta}_j$ ,  $\min_{j \in S^c} w_j = 1/\max_{j \in S^c} \hat{\beta}_j$  and re-arranging yields the condition

$$\min_{j \in S} \hat{\beta}_j > 3 \max(\varrho_{S,w}, 1)(1 + \iota_S) \max_{j \in S^c} \hat{\beta}_j.$$

Combining paragraphs (A), (B) and (C), we conclude that under the stated

conditions, I) and II) in (S4.3) hold so that  $S(\widehat{\beta}_\lambda^w) = S(\bar{\beta}) = S(\beta^*)$ . This completes the proof.

## S5 Proof of Proposition 4

The optimization problem under consideration is equivalent to the following one:

$$\min_{\beta \in \Delta^n} \left( \frac{1}{n} - \lambda \right) \|\beta\|_2^2 - \frac{2}{n} \mathbf{Z}^\top \beta. \quad (\text{S5.1})$$

For  $\lambda \geq 1/n$ , the objective becomes concave. If  $\lambda > 1/n$ , the objective is strictly concave and the unique minimum is attained at one of the vertices  $\{e_i\}_{i=1}^n$  of  $\Delta^n$ . Specifically, the minimum is attained for any  $e_i$  s.t.  $\langle \mathbf{Z}, e_i \rangle = z_i = \max_{1 \leq k \leq n} z_k$ . Since we have assumed that  $z_{(1)} > \dots > z_{(n)}$ , such vector is unique. If  $\lambda = 1/n$ , we have

$$\widehat{\beta}_\lambda^{\ell_2} \in \text{conv} \left\{ e_i : z_i = \max_{1 \leq k \leq n} z_k \right\}.$$

By the same argument as above, that convex hull equals the unique vector  $e_i$  s.t.  $z_i = \max_{1 \leq k \leq n} z_k$ .

For  $0 \leq \lambda < 1/n$ , the problem becomes strictly convex. With  $\gamma = 1 - n\lambda$ , (S5.1) is equivalent to

$$\min_{\beta \in \Delta^n} \gamma \|\beta\|_2^2 - 2\mathbf{Z}^\top \beta.$$



Re-arranging terms, this can be seen to be equivalent to

$$\min_{\beta \in \Delta^n} \|\beta - \mathbf{Z}/\gamma\|_2^2,$$

i.e.,  $\widehat{\beta}_\lambda^{\ell_2} = \Pi_{\Delta^n}(\mathbf{Z}/\gamma)$ , with  $\Pi_{\Delta^n}$  denoting the Euclidean projection onto  $\Delta^n$ .

Suppose that the realizations  $\mathbf{z} = (z_i)_{i=1}^n$  are arranged such that

$$z_1 = \beta_1^* + \varepsilon_1 > z_2 = \beta_2^* + \varepsilon_2 > \dots > z_s = \beta_s^* + \varepsilon_s > z_{s+1} = \varepsilon_{s+1} > \dots > z_p = \varepsilon_p.$$

Under the event  $\{b_{\min}^* = \min_{i \in S(\beta^*)} |\beta_i^*| \geq 2 \max_{1 \leq i \leq n} |\varepsilon_i|\}$ , this can be assumed without loss of generality. The projection of  $\mathbf{Z}/\gamma$  onto  $\Delta^n$  can then be expressed as (cf. Kyriallidis et al. (2013))

$$(\Pi_{\Delta^n}(\mathbf{Z}/\gamma))_i = \max\{z_i/\gamma - \tau, 0\}, \quad \text{where } \tau = \frac{1}{q} \left( \sum_{i=1}^q (z_i/\gamma) - 1 \right),$$

and

$$q = \max \left\{ k : (z_k/\gamma) > \frac{1}{k} \left( \sum_{i=1}^k (z_i/\gamma) - 1 \right) \right\}.$$

In order to establish that  $S(\widehat{\beta}_\lambda^{\ell_2}) = S(\beta^*)$ , it remains to be shown that under the given conditions on  $b_{\min}^*$  and  $\lambda$  respectively  $\gamma$ , the following properties

(a) and (b) hold true:

$$\begin{aligned}
 (a) \quad & \frac{\beta_s^* + \varepsilon_s}{\gamma} > \frac{1}{\gamma} \frac{\beta_1^* + \dots + \beta_s^* - \gamma}{s} + \frac{1}{\gamma} \frac{\varepsilon_1 + \dots + \varepsilon_s}{s} \\
 \iff & \frac{\beta_s^* + \varepsilon_s}{\gamma} > \frac{1}{s} \frac{1 - \gamma}{\gamma} + \frac{1}{\gamma} \frac{\varepsilon_1 + \dots + \varepsilon_s}{s} \\
 \iff & \beta_s^* > \frac{1}{s} (\{1 - \gamma\} - \{\varepsilon_1 + \dots + \varepsilon_s - s\varepsilon_s\}).
 \end{aligned}$$

$$(b) \quad \frac{\varepsilon_{s+1}}{\gamma} < \frac{1}{\gamma} \frac{1 - \gamma}{s + 1} + \frac{1}{\gamma} \frac{\varepsilon_1 + \dots + \varepsilon_s + \varepsilon_{s+1}}{s + 1}.$$

Re-arranging (b), we find that

$$n\lambda = (1 - \gamma) > s\varepsilon_{s+1} - (\varepsilon_1 + \dots + \varepsilon_s),$$

which is implied by

$$n\lambda > 2s \max_{1 \leq i \leq n} |\varepsilon_i|.$$

Likewise, the inequality in (a) holds as long as

$$\beta_s^* > \frac{n\lambda}{s} + 2 \max_i |\varepsilon_i|.$$

This concludes the proof.

## S6 Proof of Proposition 5

We provide a proof for problem (4.3) restated in (S6.1) below; the proof for problem (4.4) follows similarly. Consider the optimization problem

$$\min_{\beta \in \Delta^p} R_n(\beta) - \lambda \|\beta\|_2^2. \tag{S6.1}$$

The subproblem solved in each iteration in the case of (S6.1) is given by

$$\min_{\beta \in \Delta^p} R_n(\beta) - 2\lambda \langle \beta^k, \beta - \beta^k \rangle \quad (\text{S6.2})$$

First note that the constraint sets of (S6.1) and (S6.2) are compact and the objectives are continuous. Thus, by Weierstrass' theorem, these problems have a minimizer, and the minima are finite.

The current iterate  $\beta^k$  is always feasible for (S6.2). Hence the optimal value of (S6.2) is either  $R_n(\beta^k)$  (in which case the algorithm terminates) or strictly smaller than  $R_n(\beta^k)$ ,

$$R_n(\beta^{k+1}) - 2\lambda \langle \beta^k, \beta^{k+1} - \beta^k \rangle < R_n(\beta^k). \quad (\text{S6.3})$$

On the other hand, by convexity of  $\lambda \|\beta\|_2^2$ , we have

$$\begin{aligned} f(\beta^{k+1}) &= R_n(\beta^{k+1}) - \lambda \|\beta^{k+1}\|_2^2 \leq R_n(\beta^{k+1}) - \lambda \|\beta^k\|_2^2 - 2\lambda \langle \beta^k, \beta^{k+1} - \beta^k \rangle \\ &\stackrel{(\text{S6.3})}{<} R_n(\beta^k) - \lambda \|\beta^k\|_2^2 \\ &= f(\beta^k). \end{aligned}$$

This establishes the strict monotonicity of the iterates in terms of the objective  $f$  of the original problem (S6.1) until convergence. It is clear that all the elements of the sequence  $\{\beta^k\}$  are feasible for (S6.1) and satisfy  $f^* \leq f(\beta^k)$ ,  $k \geq 0$ , where  $f^*$  is the global minimum of (S6.1). Since  $\{f(\beta^k)\}$  is a strictly decreasing sequence bounded below by a finite  $f^*$ , the

sequence converges to a limit

$$\bar{f} = \lim_{k \rightarrow \infty} f(\beta^k).$$

Since all the elements of the sequence  $\{\beta^k\}$  are contained in  $\Delta^p$ , a compact set, there exists a subsequence  $\{\beta^{k_i}\}$  converging to an element  $\bar{\beta} \in \Delta^p$ . The sequence  $\{f(\beta^{k_i})\}$  is a subsequence of  $\{f(\beta^k)\}$  that is shown to converge to the limit  $\bar{f}$ ; hence the subsequence  $\{f(\beta^{k_i})\}$  also converges to the same limit

$$\lim_{k \rightarrow \infty} f(\beta^{k_i}) = \bar{f}.$$

Let us define  $\phi_{\bar{\beta}}(\beta) = R_n(\beta) - 2\lambda \langle \bar{\beta}, \beta - \bar{\beta} \rangle$ . We now argue that  $\bar{\beta} \in \operatorname{argmin}_{\beta \in \Delta^p} \phi_{\bar{\beta}}(\beta)$ . To see this note that  $\bar{\beta}$  is feasible for this problem and hence  $\min_{\beta \in \Delta^p} \phi_{\bar{\beta}}(\beta) \leq f(\bar{\beta}) = \bar{f}$ . Assume for the sake of contradiction that a minimizer  $\check{\beta}$  of this problem has a strictly smaller objective,

$$\phi_{\bar{\beta}}(\check{\beta}) = R_n(\check{\beta}) - 2\lambda \langle \bar{\beta}, \check{\beta} - \bar{\beta} \rangle < \bar{f}.$$

Similar to the argument above regarding strict descent, we can show that

$$f(\check{\beta}) < \bar{f},$$

which contradicts the fact that the sequence  $\{f(\beta^k)\}$  converges to the limit  $\bar{f}$ . Thus, we must have,

$$\bar{\beta} \in \operatorname{argmin}_{\beta \in \Delta^p} R_n(\beta) - 2\lambda \langle \bar{\beta}, \beta - \bar{\beta} \rangle.$$

The first-order optimality condition for  $\bar{\beta}$  then implies

$$-\nabla R_n(\bar{\beta}) + 2\lambda\bar{\beta} \in N_{\Delta^p}(\bar{\beta}),$$

where  $N_{\Delta^p}(\bar{\beta})$  is the normal cone of  $\Delta^p$  at  $\bar{\beta}$  (see, e.g., Rockafellar and Wets (2004) for a definition). Note that this is exactly the first-order optimality condition for the original problem (S6.1). Finally note that the argument is true for any subsequence  $\{\beta^{k_i}\}$  and hence each of such subsequences and consequently the original sequence  $\{\beta^k\}$  converge to the same limit  $\bar{\beta}$ , which has been shown to satisfy the required optimality condition.

## S7 Proof of Proposition 6

Before providing a proof of Proposition 5, we first provide a precise definition of the linear spaces  $\mathbb{T}(B)$ ,  $B \in \mathbf{B}_0^m(r) \subset \mathbb{H}^m$ .

**Definition 1.** Let  $B \in \mathbf{B}_0^m(r)$  have the spectral decomposition  $B = U\Lambda U^H$ , where

$$U = \begin{bmatrix} U_{\parallel} & U_{\perp} \\ m \times r & m \times (m-r) \end{bmatrix} \begin{bmatrix} \Lambda_r & 0_{r \times (m-r)} \\ 0_{(m-r) \times r} & 0_{(m-r) \times (m-r)} \end{bmatrix}$$

for  $\Lambda_r$  real and diagonal. We then define

$$\mathbb{T}(B) = \{M \in \mathbb{H}^m : M = U_{\parallel}\Gamma + \Gamma^H U_{\parallel}^H, \quad \Gamma \in \mathbb{C}^{r \times m}\}.$$

It is immediate from the definition of  $\mathbb{T}(B)$  that its orthogonal complement is given by

$$\mathbb{T}(B)^\perp = \{M \in \mathbb{H}^m : M = U_\perp A U_\perp^H, \quad A \in \mathbb{H}^{m-r}\}.$$

We first show that  $\widehat{\Phi} = \widehat{B} - B^* \in \mathcal{K}^\Delta(r)$ , where we recall that

$$\mathcal{K}^\Delta(r) = \{\Phi \in \mathbb{H}^m : \exists B \in \mathbf{B}_0^m(r) \text{ s.t.}$$

$$\text{tr}(\Pi_{\mathbb{T}(B)^\perp}(\Phi)) = -\text{tr}(\Pi_{\mathbb{T}(B)}\Phi) \text{ and } \Pi_{\mathbb{T}(B)^\perp}(\Phi) \succeq 0\}.$$

Define the shortcuts  $\widehat{\Phi}_\mathbb{T} = \Pi_{\mathbb{T}(B^*)}\widehat{\Phi}$  and  $\widehat{\Phi}_{\mathbb{T}^\perp} = \Pi_{\mathbb{T}(B^*)^\perp}\widehat{\Phi}$ . Since  $\widehat{B}$  is feasible, it must hold that  $\text{tr}(\widehat{\Phi}) = 0$  and thus  $\text{tr}(\widehat{\Phi}_{\mathbb{T}^\perp}) = -\text{tr}(\widehat{\Phi}_\mathbb{T})$ . Since  $\widehat{B}$  must also be positive definite, it must hold that  $\text{tr}(\widehat{B}W) \geq 0$  for all  $W \in \mathbb{T}(B^*)^\perp$ ,  $W \succeq 0$ . We have

$$\text{tr}(\widehat{B}W) = \text{tr}((B^* + \widehat{\Phi})W) = \text{tr}(\widehat{\Phi}_{\mathbb{T}^\perp}W) \quad \forall W \in \mathbb{T}(B^*)^\perp,$$

since  $B^* \in \mathbb{T}(B^*)$ . We conclude that  $\text{tr}(\widehat{\Phi}_{\mathbb{T}^\perp}W) \geq 0$  for all  $W \in \mathbb{T}(B^*)^\perp$ ,

$W \succeq 0$ , and thus  $\widehat{\Phi}_{\mathbb{T}^\perp} \succeq 0$ . Altogether, we have shown that  $\widehat{\Phi} \in \mathcal{K}^\Delta(r)$ .

Since  $\widehat{B}$  is a minimizer, we have

$$\frac{1}{n} \|\mathbf{Y} - \mathcal{X}(\widehat{B})\|_2^2 \leq \frac{1}{n} \|\mathbf{Y} - \mathcal{X}(B^*)\|_2^2$$

After re-arranging terms, we obtain

$$\begin{aligned}
 \frac{1}{n} \|\mathcal{X}(B^* - \widehat{B})\|_2^2 &\leq \frac{2}{n} \langle \varepsilon, \mathcal{X}(\widehat{B} - B^*) \rangle \\
 &= \frac{2}{n} \langle \mathcal{X}^*(\varepsilon), \widehat{B} - B^* \rangle \\
 &\leq 2 \|\mathcal{X}^*(\varepsilon)/n\|_\infty \|\widehat{B} - B^*\|_1 \\
 &= \lambda_* \|\widehat{B} - B^*\|_1.
 \end{aligned}$$

where  $\mathcal{X}^*$  is the adjoint of  $\mathcal{X}$ . By  $\Delta$ -RSC, we now have

$$\frac{1}{n} \|\mathcal{X}(B^* - \widehat{B})\|_2^2 \geq \kappa \|B^* - \widehat{B}\|_2^2.$$

Combining this with the preceding upper bound, we hence obtain

$$\begin{aligned}
 \|\widehat{B} - B^*\|_2^2 &\leq \frac{\lambda_*^2}{\kappa^2} \left( \frac{\|\widehat{B} - B^*\|_1}{\|\widehat{B} - B^*\|_2} \right)^2 \leq \frac{8r\lambda_*^2}{\kappa^2}, \\
 \|\widehat{B} - B^*\|_1 &\leq \frac{\lambda_*}{\kappa} \left( \frac{\|\widehat{B} - B^*\|_1}{\|\widehat{B} - B^*\|_2} \right)^2 \leq \frac{8r\lambda_*}{\kappa},
 \end{aligned}$$

The rightmost inequalities follow from the fact that  $\widehat{B} - B^* = \widehat{\Phi} \in \mathcal{K}^\Delta(r)$

and hence  $\|\widehat{\Phi}_{\mathbb{T}^\perp}\|_1 \leq \|\widehat{\Phi}_{\mathbb{T}}\|_1$  so that

$$\begin{aligned}
 \|\widehat{B} - B^*\|_1 &= \|\widehat{\Phi}\|_1 = \|\widehat{\Phi}_{\mathbb{T}}\|_1 + \|\widehat{\Phi}_{\mathbb{T}^\perp}\|_1 \\
 &\leq 2\|\widehat{\Phi}_{\mathbb{T}}\|_1 \\
 &\leq 2\sqrt{2r} \|\widehat{\Phi}_{\mathbb{T}}\|_2 \leq 2\sqrt{2r} \|\widehat{B} - B^*\|_2,
 \end{aligned}$$

where for the third inequality, we have used that  $\|M\|_0 \leq 2r$  for all  $M \in$

$\mathbb{T}(B^*)$ .

The bound for  $\tilde{B}_\lambda$  can be established by combining the proof scheme used for  $\tilde{\beta}_\lambda$  with the scheme used for  $\hat{B}$  and is thus omitted.

## S8 Proof of Proposition 7

We start by expanding the objective function of the optimization problem under consideration. Define  $\mathbb{S}^m := \mathbb{H}^m \cap \mathbb{R}^{m \times m}$  which is a subspace of  $\mathbb{H}^m$  that is isometrically isomorphic (w.r.t. the standard inner product) to  $\mathbb{R}^{\delta_m}$ ,  $\delta_m = m(m+1)/2$  under the isometry  $\mathcal{X}$  (5.22). Therefore,

$$\begin{aligned} \frac{1}{n} \|\mathbf{Y} - \mathcal{X}(B)\|_2^2 &= \frac{1}{n} \|\mathcal{X}^*(\mathbf{Y}) - B\|_2^2 \\ &= \frac{1}{n} \|B^* + E - B\|_2^2, \quad E := \mathcal{X}^*(\varepsilon), \\ &= \frac{1}{n} \|\Upsilon - B\|_2^2, \quad \Upsilon := B^* + E. \end{aligned} \quad (\text{S8.1})$$

It follows directly from the definition of  $\mathcal{X}^*$  that the symmetric random matrix  $E = (\varepsilon_{jk})_{1 \leq j, k \leq m}$  is distributed according to the Gaussian orthogonal ensemble (GOE, see e.g. Tao (2012)), i.e.,  $E \sim \text{GOE}(m)$ , where

$$\begin{aligned} \text{GOE}(m) &= \{X = (x_{jk})_{1 \leq j, k \leq m}, \{x_{jj}\}_{j=1}^m \stackrel{\text{i.i.d.}}{\sim} N(0, 1/m), \\ &\quad \{x_{jk} = x_{kj}\}_{1 \leq j < k \leq m} \stackrel{\text{i.i.d.}}{\sim} N(0, 1/2m)\}. \end{aligned}$$

In virtue of (S8.1), we have

$$\min_{B \in \Delta^m} \frac{1}{n} \|\mathbf{Y} - \mathcal{X}(B)\|_2^2 = \min_{B \in \Delta^m} \left\{ (1/n - \lambda) \|B\|_2^2 - \frac{2}{n} \langle \Upsilon, B \rangle \right\} + \frac{1}{n} \|\Upsilon\|_2^2.$$



At this point, the proof parallels the proof of Proposition 4. We see that for  $\lambda \geq 1/n$ ,  $\widehat{B}_\lambda^{\ell_2} = u_1 u_1^\top$ , where  $u_1$  is the eigenvector of  $\Upsilon$  corresponding to its largest eigenvalue. This follows from the duality of the Schatten  $\ell_1/\ell_\infty$  norms and the fact that for all feasible  $B$ , it holds that  $\|B\|_2^2 \leq \|B\|_1^2 = 1$  with equality if and only if  $B$  has rank one. Conversely, if  $0 \leq \lambda < 1/n$ , we define  $\gamma := 1 - n\lambda > 0$  and deduce that the optimization problem in the previous display is equivalent to  $\min_{B \in \Delta^m} \|\Upsilon/\gamma - B\|_2^2$  with minimizer  $\widehat{B}_\lambda^{\ell_2} = U \text{diag}(\{\widehat{\phi}_j\}_{j=1}^m) U^\top$ , where  $\widehat{\phi} = \Pi_{\Delta^m}(v/\gamma)$  with  $v = (v_j)_{j=1}^m$  denoting the eigenvalues of  $\Upsilon$  (in decreasing order) corresponding to the eigenvectors in  $U$ . We now prove the last claim of the proposition, combining the proof of Proposition 4 for the vector case with concentration results by Peng (2012) for the spectrum of the random matrix  $\Upsilon = B^* + E$ , which are here rephrased as follows. Define

$$\widetilde{\phi}_j^* = \begin{cases} \phi_j^* + \frac{\sigma^2}{\phi_j^*} & \text{if } \sigma < \phi_j^* \leq 1 \\ 2\sigma & \text{if } 0 \leq \phi_j^* \leq \sigma, \quad j = 1, \dots, m, \end{cases}$$

where we recall that the  $\{\phi_j^*\}_{j=1}^m$  denote the ordered eigenvalues of  $B^*$  and  $\sigma^2$  is the variance of the noise (up to a scaling factor of  $1/m$ ). We then have

$$\mathbf{P}(v_j \geq \widetilde{\phi}_j^* + t) \leq C_1 \exp(-c_1 m t^2 / \sigma^2), \quad j = 1, \dots, m.$$

Furthermore, let  $r_0$  denote the number of eigenvalues of  $B^*$  that are larger than  $\sigma$ . Then, there is a constant  $c_0 > 0$  so that if  $r \leq c_0 m$ , it holds that

$$\mathbf{P}(v_j \leq \tilde{\phi}_j^* - t - 2\sigma) \leq \exp(-c_2 m / \sigma^2) + C_2' \exp(-c_2' m t^2 / \sigma^2), \quad j = 1, \dots, r_0,$$

where  $c_1, c_2, C_1, C_2, C_2'$  are positive constants.

It needs to be shown that for a suitable choice of  $\lambda$  and for  $\phi_r^*$  large enough, it holds that  $\|\widehat{B}_\lambda^{\ell_2}\|_0 = \|B^*\|_0 = r$  with high probability as specified in the proposition. This is the case if and only if  $\widehat{\phi} = \Pi_{\Delta^m}(v/\gamma)$  has precisely  $r$  non-zero entries.

a)  $\|\widehat{\phi}\|_0 \geq r$ :

It follows from the proof in the vector case that a) is satisfied if

$$\frac{v_r}{\gamma} > \frac{v_1 + \dots + v_r - \gamma}{r\gamma}$$

Write  $\xi_j = v_j - \tilde{\phi}_j^*$ ,  $b_j = \tilde{\phi}_j^* - \phi_j^*$ ,  $j = 1, \dots, m$ , and  $\bar{\xi} = \max_{1 \leq j \leq m} \xi_j$ ,  $\underline{\xi} = \min_{1 \leq j \leq r_0} \xi_j$ . Then the above condition can equivalently be expressed

as

$$\begin{aligned} v_r &> \frac{1}{r} \left\{ \sum_{j=1}^r (\tilde{\phi}_j^* + \xi_j) - \gamma \right\} \\ &= \frac{1}{r} \left\{ \sum_{j=1}^r (b_j + \xi_j) + (1 - \gamma) \right\}, \quad \text{since } \sum_{j=1}^r \phi_j^* = 1 \\ &= \frac{1}{r} \sum_{j=1}^r (b_j + \xi_j) + \frac{n\lambda}{r} \end{aligned}$$

As  $\phi_j^* \geq 5\sigma$  for  $j = 1, \dots, r$  by assumption, we have  $r = r_0$  and

$$\frac{1}{r} \sum_{j=1}^r (b_j + \xi_j) \leq \sigma + \bar{\xi}.$$

Since  $v_r \geq \phi_r^* + \underline{\xi}$ , we obtain the sufficient condition

$$(A) \quad \phi_r^* > -\underline{\xi} + \sigma + \bar{\xi} + \frac{n\lambda}{r}.$$

b)  $\|\widehat{\phi}\|_0 \leq r$

In analogy to a), we start with the condition

$$\frac{v_{r+1}}{\gamma} < \frac{v_1 + \dots + v_r + v_{r+1} - \gamma}{(r+1)\gamma}$$

After canceling  $\gamma$  on both sides, we lower bound the right hand side as follows:

$$\frac{v_1 + \dots + v_r + v_{r+1} - \gamma}{r+1} \geq \frac{(1-\gamma) + v_{r+1} + r\underline{\xi}}{r+1}.$$

Back-substituting this lower bound, we obtain the following sufficient condition

$$(B) \quad \lambda > \frac{r}{n}(v_{r+1} - \underline{\xi}).$$

Consider the following two events:

$$E_1 : \{\bar{\xi} > \sigma\}, \quad E_2 : \{\underline{\xi} < -3\sigma\}$$

The concentration results stated above yield that  $\mathbf{P}(E_1 \cup E_2) \leq C \exp(-cm)$

for constants  $c, C > 0$ . Note that conditional on the complement of  $E_1 \cup E_2$ ,

## REFERENCES

---

$v_{r+1} \leq 3\sigma$  so that condition (B) is fulfilled as long as  $\lambda > 6\sigma r/n$ . Likewise, condition (A) is fulfilled as long as  $\phi_r^* > 5\sigma + n\lambda/r$ .

### References

- Brodie, J., I. Daubechies, C. D. Mol, D. Giannone, and I. Loris (2009). Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences* 106, 12267–12272.
- Bunea, F., A. Tsybakov, M. Wegkamp, and A. Barbu (2010). SPADES and mixture models. *The Annals of Statistics* 38, 2525–2558.
- Donoho, D. and J. Tanner (2005). Neighbourliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Science* 102, 9452–9457.
- Gross, D., Y.-K. Liu, S. Flammia, S. Becker, and J. Eisert (2010). Quantum State Tomography via Compressed Sensing. *Physical Review Letters* 105, 150401–15404.
- Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics* 5, 1161–1183.
- Koltchinskii, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics* 39, 2936–2973.
- Kyriillidis, A., S. Becker, V. Cevher, and C.Koch (2013). Sparse projections onto the simplex. In *International Conference on Machine Learning (ICML)*, Volume 28 of *JMLR W&CP*, pp. 235–243.
- Kyriillidis, A. and V. Cevher (2011). Recipes on Hard Thresholding Methods. In *International*

- Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 353–356.
- Liu, Y. (2011). Universal low-rank matrix recovery from Pauli measurements. In *Advances in Neural Information Processing Systems*, Volume 24, pp. 1638–1646.
- Negahban, S., P. Ravikumar, M. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of  $M$ -estimators with decomposable regularizers. *Statistical Science* 27, 538–557.
- Negahban, S. and M. Wainwright (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics* 39, 1069–1097.
- Peng, M. (2012). Eigenvalues of Deformed Random Matrices. arXiv:1205.0572.
- Pilanci, M., L. E. Ghaoui, and V. Chandrasekaran (2012). Recovery of Sparse Probability Measures via Convex Programming. In *Advances in Neural Information Processing Systems (NIPS)*, Volume 25, pp. 2420–2428.
- Recht, B., M. Fazel, and P. Parillo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review* 52, 471–501.
- Rockafellar, T. and R. Wets (2004). *Variational Analysis*. Springer.
- Slawski, M. and M. Hein (2013). Non-negative least squares for high-dimensional linear models: consistency and sparse recovery without regularization. *The Electronic Journal of Statistics* 7, 3004–3056.

## REFERENCES

---

Tao, T. (2012). *Topics in Random Matrix Theory*. American Mathematical Society.

van de Geer, S., P. Bühlmann, and S. Zhou (2013). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electronic Journal of Statistics* 5, 688–749.