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 an 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, $i = 1, \ldots, n$. In short, $\mathbf{Y} = \boldsymbol{X}\beta^*$ with $\mathbf{Y} = (Y_i)_{i=1}^n$ and $\boldsymbol{X}$ having the $\{X_i\}_{i=1}^n$ as its rows. Identifying β^* with a probability distribution π on $\{1, \ldots, p\}$, we may think of the problem as recovering π 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 β^* 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 $R_n(\beta) = \|\mathbf{Y} - \boldsymbol{X}\beta\|_2^2/n$:

'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': ℓ_2-norm maximization (4.4) with $\lambda = 0$, i.e., over

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

'Pilanci': The method of Pilanci, Ghaoui, and Chandrasekaran (2012) that maximizes the ℓ_∞-norm over (S1.1).

'weighted L1': Weighted ℓ_1-norm minimization (cf. §3) over (S1.1).

'IHT': Iterative hard threshold under simplex constraints (Kyrillidis et al., 2013). Regarding the step size used for gradient projection, we use the method in Kyrillidis 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 ℓ_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 $n/p \to c \in (0,1)$ and $s/n \to c' \in (0,1)$. 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'

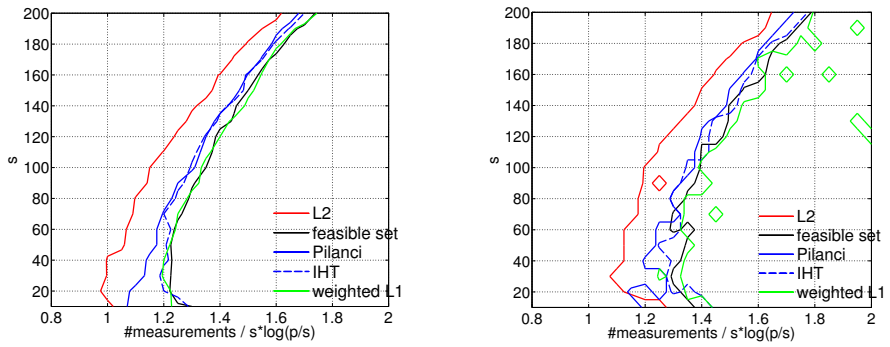


Figure 1: Contour plots of the empirical relative frequencies of exact recovery in dependency of the number of measurements (horizontal axis) and s (vertical axis). The left and right plot show the contour levels .75 and .99, respectively. Note that the smaller the area "left" to and "above" the curve,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

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

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 β^* 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 ℓ_2-regularization (4.3). For the parameter λ, we consider a grid Λ of 100 logarithmically spaced points from 0.01 to $\phi_{\max}(\boldsymbol{X}^\top\boldsymbol{X}/n)$, the maximum eigenvalue of $\boldsymbol{X}^\top\boldsymbol{X}/n$. Note that for $\lambda \geq \phi_{\max}(\boldsymbol{X}^\top\boldsymbol{X}/n)$, the optimization problem (4.3) becomes concave and the minimizer must consequently be a vertex of Δ^p, i.e., the solution is maximally sparse at this point, and it hence does not make sense to consider even larger values of λ. When computing the solutions $\{\widehat{\beta}^{\ell_2}_\lambda,\, \lambda \in \Lambda\}$, we use a homotopy-type scheme in which for each $\lambda \in \Lambda$, Algorithm 1 is initialized with the solution for the previous λ, using the output $\widehat{\beta}$ of 'ERM' as initialization for the smallest value of λ.

'L2-D': ℓ_2-norm maximization (4.4) over $\mathcal{D}(C\lambda_0)$ with λ_0 being the noise level defined above and $C \in \{0.5, 0.55, \ldots, 2\}$. 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 ℓ_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

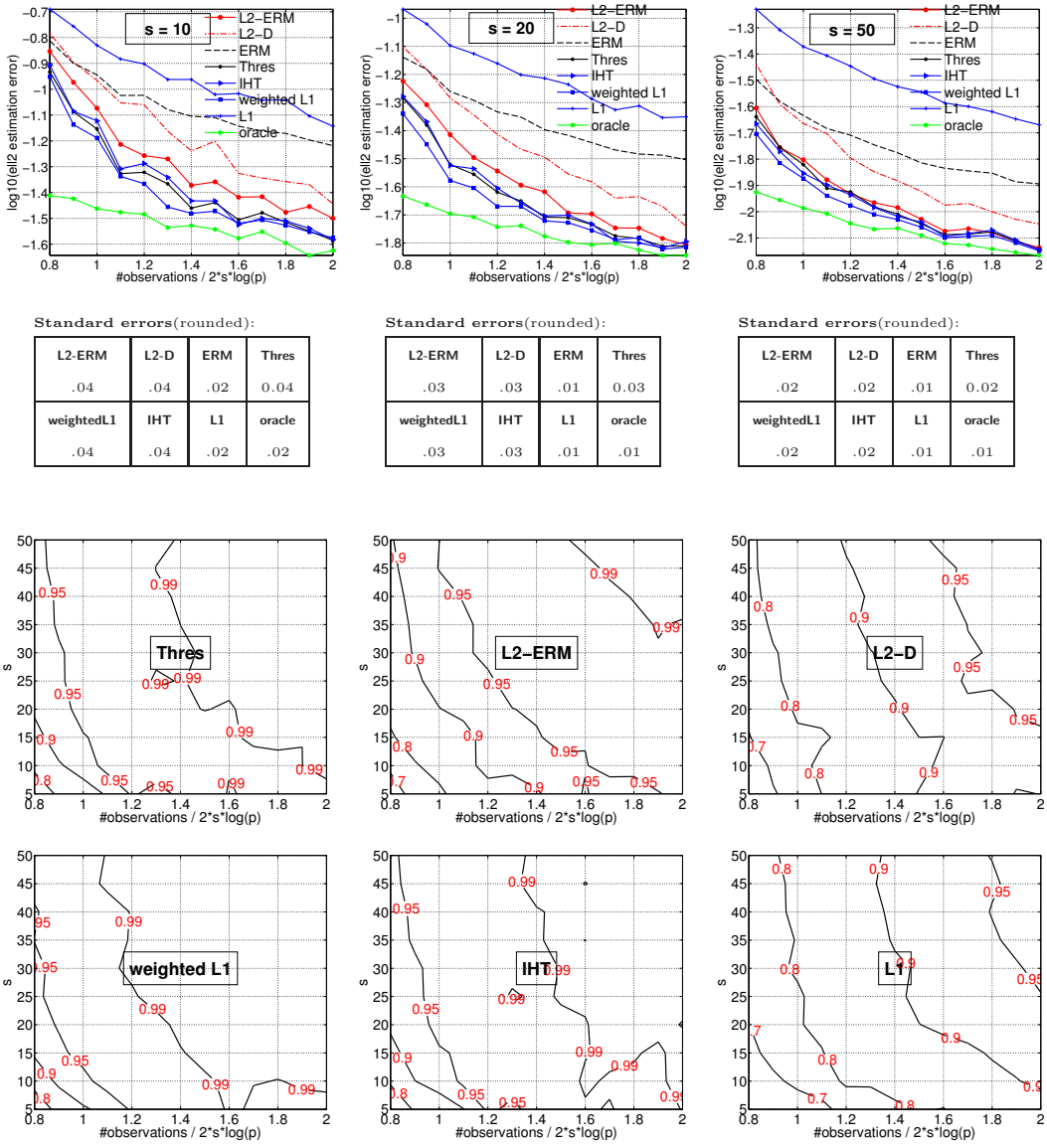


Figure 2: Upper panel: Average estimation errors $\|\widehat{\theta}-\beta^*\|_2$ ($\log_{10}$ scale) in dependence of n over 50 trials for selected values of s. Here, $\widehat{\theta}$ is a placeholder for any of the estimators under consideration. Middle and Lower panel: contour plots of the average 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.

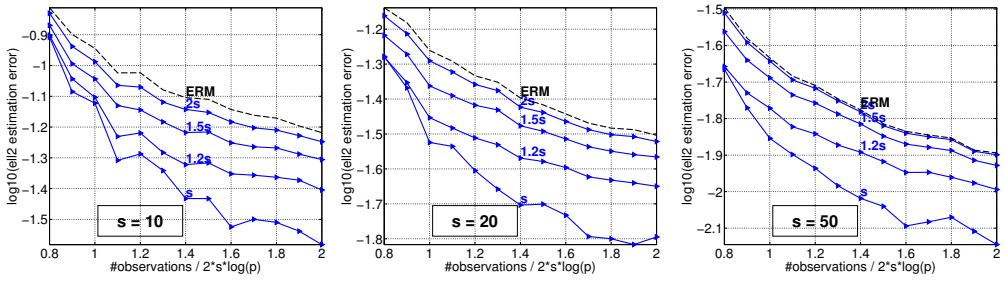


Figure 3: Sensitivity of 'IHT' w.r.t. the choice of s. The plots display error curves of IHT run with the correct value of s as appearing in Figure 2 as well with overspecification of s by the factors $1.2, 1.5, 2$. The drop in performance is substantial: for $2s$, the improvement over ERM (here used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

higher ℓ_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 ℓ_1-regularization for the most part outperform the more sophisticated methods. Among the two methods based on negative ℓ_2-regularization, 'L2-ERM' achieves better performance than 'L2-D'. We also investigate success in support recovery by comparing $S(\widehat{\theta})$ and $S(\beta^*)$, where $\widehat{\theta}$ represents any of the considered estimators, by means of Matthew's correlation coefficient (MCC) defined by

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

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 ℓ_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 ϕ_{β^*}, 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, $j = 1, \ldots, p$. As an example, one might consider p_0 locations and K different standard deviations per location so that $p = p_0 K$, i.e., $\theta_{(k-1)p_0+l} = (\mu_l, \sigma_k)$, $k = 1, \ldots, K$, and $l = 1, \ldots, p_0$. 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 β 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

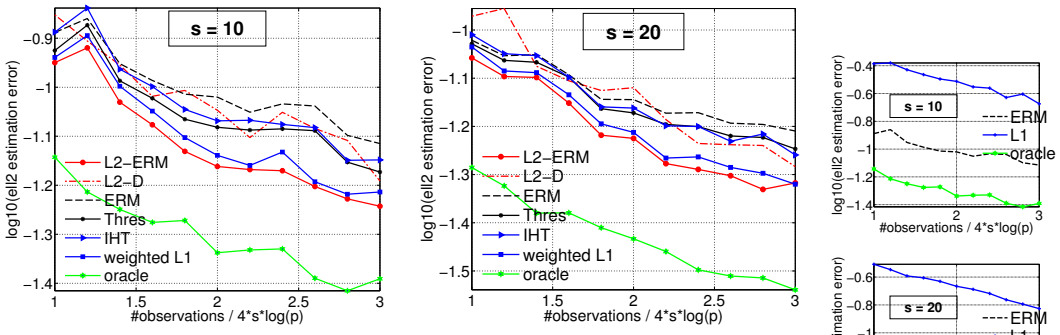
$$R_n(\beta) = \beta^\top Q\beta - 2c^\top\beta, \quad c = (\textstyle\sum_{i=1}^n \phi_j(Z_i)/n)_{j=1}^p\,,$$

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 μ_1 randomly from $[0, \delta]$, μ_2 from $[\mu_1 + \delta, \mu_1 + 2\delta]$ etc. where δ 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), $b^*_{\min}$ (ϱ 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 ℓ_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).



Standard errors(rounded):

L2-ERM	L2-D	ERM	Thres
.03	.03	.02	0.02
weightedL1	**IHT**	**L1**	**oracle**
.03	.03	.02	.02

Standard errors(rounded):

L2-ERM	L2-D	ERM	Thres
.02	.02	.02	0.02
weightedL1	**IHT**	**L1**	**oracle**
.02	.02	.01	.01

Figure 4: Average estimation errors $\|\widehat{\theta} - \beta^*\|_2$ 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` 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 $\widehat{\mu}$ from the first four years (208 weeks). Likewise, the covariance of the returns is estimated as the sample covariance $\widehat{\Sigma}$ of the returns of the first four years. Given $\widehat{\mu}$ and $\widehat{\Sigma}$, portfolio selection (without short positions) is based on the optimization problem

$$\min_{\beta \in \Delta^p} \beta^\top \widehat{\Sigma} \beta - \tau \widehat{\mu}^\top \beta \tag{S1.3}$$

where $\tau \in [0, \tau_{\max}]$ is a parameter controlling the trade-off between return and variance of the portfolio. Assuming that $\widehat{\mu}$ 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 $\widehat{\mu}$. 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 $\widehat{\mu}$ and $\widehat{\Sigma}$.

Results. Figure 5 displays the Sharpe ratios of the portfolios returned by these approaches in dependency of the ℓ_2-norms of the solutions correspond-

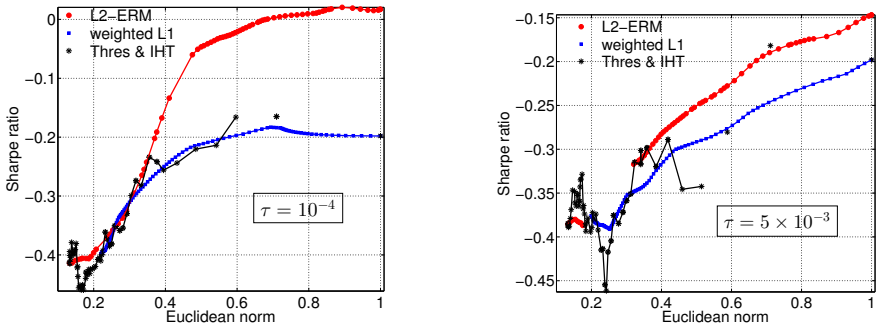


Figure 5: Sharpe ratios of the portfolios selected by 'L2-ERM', 'weighted L1', 'Thres' and 'IHT' on the hold-out portion of the NASDAQ data set in dependency of different choices for the regularization parameter/sparsity level (to allow for joint display, we use the ℓ_2-norm as measure of sparsity on 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 τ 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 ℓ_2-norms for both values of τ.

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 $\{P_{q,1}, \ldots, P_{q,m^2}\}$ 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 $\mathcal{I} \subseteq \{1, \ldots, m^2\}$, $|\mathcal{I}| = n$, 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 $B^* \in \boldsymbol{\Delta}^m$ 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

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

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 ℓ_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 $\min_{B \in \boldsymbol{\Delta}_0^m(r)} R_n(B)$ using projected gradient descent. Projections onto $\boldsymbol{\Delta}_0^m(r)$ 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 ℓ_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.

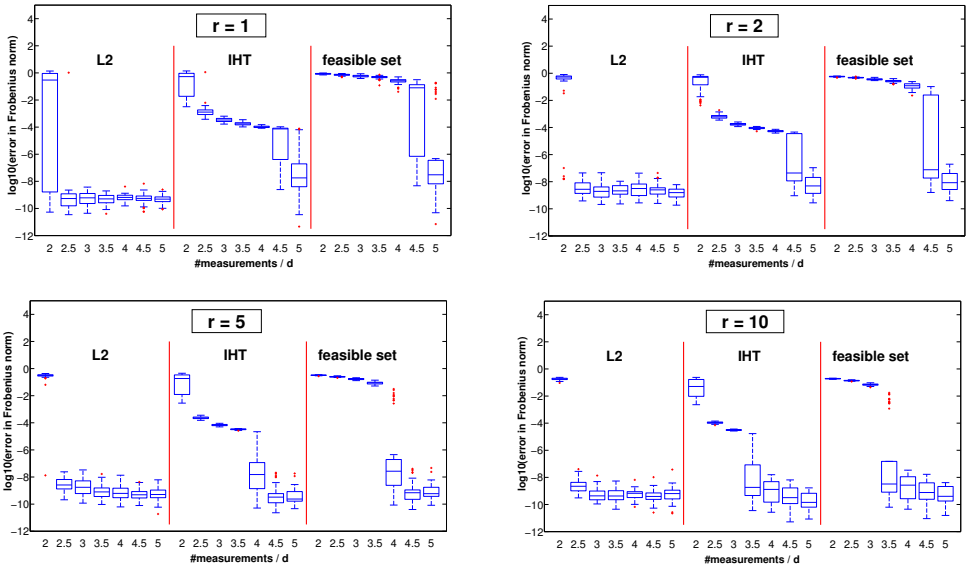


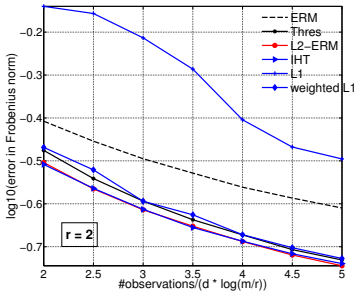
Figure 6: Boxplots of the errors $\|\widehat{\Theta} - B^*\|_2$ (50 trials) in recovering B^* with respect to the Frobenius norm ($\log_{10}$ scale) in dependence of the number of Pauli measurements (d ='degrees of freedom'). Here, $\widehat{\Theta}$ is representative for any of the three estimators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

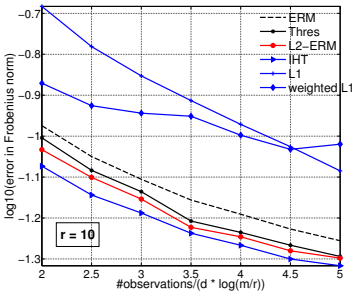
'Thres': 'ERM' and eigenvalue thresholding, outlined below Proposition 6.

'L2-ERM': Regularized ERM with negative ℓ_2-regularization (5.4). A grid search over 20 different values of the regularization parameter λ is performed analogously to the vector case.

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

'IHT': As in the noiseless case.



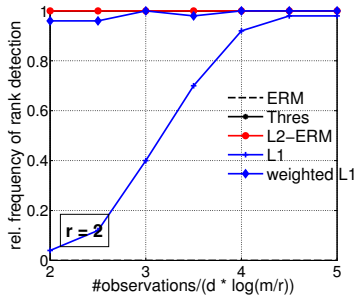


Standard errors(rounded):

L2-ERM	ERM	Thres
$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	$3 \cdot 10^{-3}$
weightedL1	**IHT**	**L1**
$1.4 \cdot 10^{-2}$	$3 \cdot 10^{-3}$	$1.7 \cdot 10^{-2}$

Standard errors(rounded):

L2-ERM	ERM	Thres
$2 \cdot 10^{-3}$	$1.2 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$
weightedL1	**IHT**	**L1**
$1.2 \cdot 10^{-2}$	$1.1 \cdot 10^{-2}$	$8 \cdot 10^{-3}$



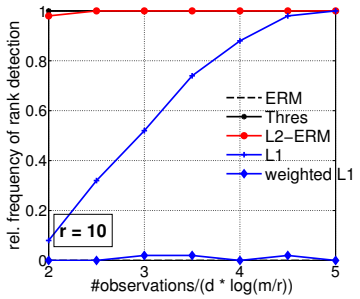


Figure 7: Bottom: Average estimation errors $\|\widehat{\Theta} - B^*\|_2$ over 50 trials ($\log_{10}$-scale) in dependence of the number of measurements ($d =$ 'degrees of freedom'). Top: Relative frequency of rank detection, i.e., of the event $\{\|\widehat{\Theta}\|_0 = \|B^*\|_0\}$; for 'IHT' this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

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

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 λ 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.

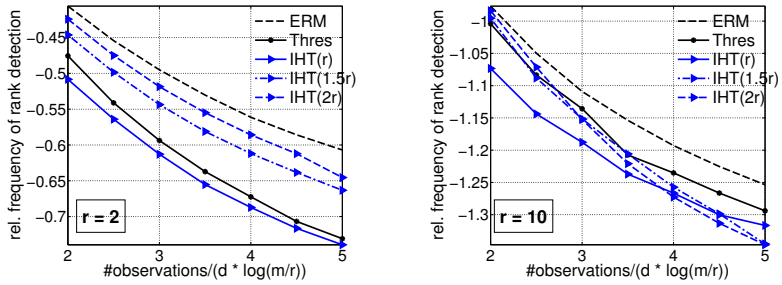


Figure 8: Sensitivity of 'IHT' w.r.t. the choice of r. The dashed-dotted and dashed lines show the average estimation errors when 'IHT' is run with $1.5r$ and $2r$, respectively. The results of 'Thres' and 'ERM' serve as reference.

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 $\widetilde{\beta}_\lambda$. Given the above bound on $R_n(\widetilde{\beta}_\lambda) - R_n(\beta^*)$, the proof can be completed by following the scheme used for $\widehat{\beta}$.

S3 Proof of Proposition 2

Invoking the Δ-RSC condition, we have

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

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

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

Combining these two bounds, we obtain that

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

This implies that

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

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

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

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 $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has the $\{X_i\}_{i=1}^n$ as it rows, and $\boldsymbol{X}_S$, $\boldsymbol{X}_{S^c}$ denote the column submatrices corresponding to S respectively S^c. Accordingly, we let $\Sigma_{SS} = \frac{1}{n}\boldsymbol{X}_S^\top\boldsymbol{X}_S$, $\Sigma_{S^cS^c} = \frac{1}{n}\boldsymbol{X}_{S^c}^\top\boldsymbol{X}_{S^c}$ and $\Sigma_{S^cS} = \frac{1}{n}\boldsymbol{X}_{S^c}^\top\boldsymbol{X}_S$. We recall that

$w = (w_j)_{j=1}^p$ with $w_j = 1/\widehat{\beta}_j$, $j = 1, \ldots, p$, 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

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

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

I) $\bar{\beta}_S \succ 0$,

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

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

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

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

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

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

$$\begin{aligned} \lambda \min_{j \in S^c} w_j &> 2\lambda \max(\varrho_{S,w}, 1)(1 + \iota_S)\|w_S\|_\infty + T'_\varepsilon, \\ T'_\varepsilon &:= \|\bar{\mu}_0(\mathbf{1} - \Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbf{1}) + \boldsymbol{X}_{S^c}^\top (I - \mathcal{P})\varepsilon/n\|_\infty, \end{aligned} \tag{S4.5}$$

with ι_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

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

Inserting $\lambda = \lambda_0 \|w_S\|_\infty$ into (S4.4) and (S4.6) with λ_0 still to be determined, we obtain the events

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

(A) Regarding T''_ε, 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 $\mathrm{tr}(\Sigma_{SS}) = s$, one shows that $\{\mathbf{1}^\top \Sigma_{SS}^{-1}\mathbf{1}\}^{-1} \leq \frac{1}{s}\max_{\|v\|_2 \leq 1} v^\top \Sigma_{SS} v \leq \frac{1}{s}\mathrm{tr}(\Sigma_{SS}) = 1$. Likewise, we note that each component of $\boldsymbol{X}_{S^c}^\top (I - \mathcal{P})\varepsilon/n$ is a Gaussian random variable with variance at

most σ^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_ε in (S4.4). First, each component of $\Sigma_{SS}^{-1}\boldsymbol{X}_S^\top\varepsilon/n$ is a Gaussian random variable with variance at most $\phi_S^{-1}\sigma^2/n$, where ϕ_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_ε 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 $\|w_S\|_\infty = 1/\min_{j\in S}\widehat{\beta}_j$, $\min_{j\in S^c} w_j = 1/\max_{j\in S^c}\widehat{\beta}_j$ and re-arranging yields the condition

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

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

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

S5 Proof of Proposition 4

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

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

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 Δ^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 $z_{(1)} > \ldots > z_{(n)}$, such vector is unique. If $\lambda = 1/n$, we have

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

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 Π_{Δ^n} denoting the Euclidean projection onto Δ^n. Suppose that the realizations $\mathbf{z} = (z_i)_{i=1}^n$ are arranged such that

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

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 Δ^n can then can be expressed as (cf. Kyrillidis et al. (2013))

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

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 λ respectively γ, the following properties

(a) and (b) hold true:

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

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

Re-arranging (b), we find that

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

which is implied by

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

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

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

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 β^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 $R_n(\beta^k)$,

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

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

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

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

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

Since all the elements of the sequence $\{\beta^k\}$ are contained in Δ^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

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

Let us define $\phi_{\bar{\beta}}(\beta) = R_n(\beta) - 2\lambda \left\langle \bar{\beta}, \beta - \bar{\beta} \right\rangle$. 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,

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

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,

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

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 Δ^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 $\mathbb{T}(B)$, $B \in \boldsymbol{B}_0^m(r) \subset \mathbb{H}^m$.

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

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

for Λ_r real and diagonal. We then define

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

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

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

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

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

since $B^* \in \mathbb{T}(B^*)$. We conclude that $\operatorname{tr}(\widehat{\Phi}_{\mathbb{T}^{\perp}}W) \geq 0$ 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 $\widehat{\Phi} \in \mathcal{K}^{\mathbf{\Delta}}(r)$. 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

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

where $\mathcal{X}^\star$ is the adjoint of $\mathcal{X}$. By $\boldsymbol{\Delta}$-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 $\widehat{B} - B^* = \widehat{\Phi} \in \mathcal{K}^{\boldsymbol{\Delta}}(r)$ 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 $\widetilde{B}_\lambda$ can be established by combining the proof scheme used for $\widetilde{\beta}_\lambda$ with the scheme used for $\widehat{B}$ 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,

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

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

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

In virtue of (S8.1), we have

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

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 Υ corresponding to its largest eigenvalue. This follows from the duality of the Schatten ℓ_1/ℓ_∞ 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 $\min_{B \in \boldsymbol{\Delta}^m} \|\Upsilon/\gamma - B\|_2^2$ with minimizer $\widehat{B}_\lambda^{\ell_2} = U \mathrm{diag}(\{\widehat{\phi}_j\}_{j=1}^m) U^\top$, where $\widehat{\phi} = \Pi_{\Delta^m}(\upsilon/\gamma)$ with $\upsilon = (\upsilon_j)_{j=1}^m$ denoting the eigenvalues of Υ (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

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

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

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

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

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

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 λ and for ϕ_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 $\widehat{\phi} = \Pi_{\Delta^m}(\upsilon/\gamma)$ 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

$$\frac{\upsilon_r}{\gamma} > \frac{\upsilon_1 + \ldots + \upsilon_r - \gamma}{r\gamma}$$

Write $\xi_j = \upsilon_j - \widetilde{\phi}_j^*$, $b_j = \widetilde{\phi}_j^* - \phi_j^*$, $j = 1, \ldots, m$, and $\overline{\xi} = \max_{1 \leq j \leq m} \xi_j$, $\underline{\xi} = \min_{1 \leq j \leq r_0} \xi_j$. Then the above condition can equivalently be expressed as

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

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

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

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

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

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

In analogy to a), we start with the condition

$$\frac{\upsilon_{r+1}}{\gamma} < \frac{\upsilon_1 + \ldots + \upsilon_r + \upsilon_{r+1} - \gamma}{(r+1)\gamma}$$

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

$$\frac{\upsilon_1 + \ldots + \upsilon_r + \upsilon_{r+1} - \gamma}{r+1} \geq \frac{(1-\gamma) + \upsilon_{r+1} + r\underline{\xi}}{r+1}.$$

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

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

Consider the following two events:

$$E_1 : \ \{\overline{\xi} > \sigma\}, \qquad E_2 : \ \{\underline{\xi} < -3\sigma\}$$

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 $E_1 \cup E_2$,

$\upsilon_{r+1} \leq 3\sigma$ 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.

Kyrillidis, 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.

Kyrillidis, 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.