Consistency of Trace Norm Minimization

Francis R. Bach   FRANCIS.BACH@MINES.ORG
INRIA - WILLOW Project-Team
Laboratoire d'Informatique de l'Ecole Normale Supérieure
CNRS/ENS/INRIA UMR 8548
45, rue d'Ulm
75230 Paris, France

Abstract

Regularization by the sum of singular values, also referred to as the *trace norm*, is a popular technique for estimating low rank rectangular matrices. In this paper, we extend some of the consistency results of the Lasso to provide necessary and sufficient conditions for rank consistency of trace norm minimization with the square loss. We also provide an adaptive version that is rank consistent even when the necessary condition for the non adaptive version is not fulfilled.

Keywords: convex optimization, singular value decomposition, trace norm, consistency

1. Introduction

In recent years, regularization by various non Euclidean norms has seen considerable interest. In particular, in the context of linear supervised learning, norms such as the ℓ_1-norm may induce sparse loading vectors, that is, loading vectors with low cardinality or ℓ_0-norm. Such regularization schemes, also known as the Lasso (Tibshirani, 1994) for least-square regression, come with efficient path following algorithms (Efron et al., 2004). Moreover, recent work has studied conditions under which such procedures consistently estimate the sparsity pattern of the loading vector (Yuan and Lin, 2007; Zhao and Yu, 2006; Zou, 2006).

When learning on rectangular matrices, the rank is a natural extension of the cardinality, and the sum of singular values, also known as the trace norm or the nuclear norm, is the natural extension of the ℓ_1-norm; indeed, as the ℓ_1-norm is the convex envelope of the ℓ_0-norm on the unit ball (i.e., the largest lower bounding convex function) (Boyd and Vandenberghe, 2003), the trace norm is the convex envelope of the rank over the unit ball of the spectral norm (Fazel et al., 2001). In practice, it leads to low rank solutions (Fazel et al., 2001; Srebro et al., 2005) and has seen recent increased interest in the context of collaborative filtering (Srebro et al., 2005), multi-task learning (Abernethy et al., 2006; Argyriou et al., 2007; Abernethy et al., 2008) or classification with multiple classes (Amit et al., 2007).

In this paper, we consider the rank consistency of trace norm regularization with the square loss, that is, if the data were actually generated by a low-rank matrix, will the matrix and its rank be consistently estimated? In Section 4, we provide necessary and sufficient conditions for the rank consistency that are extensions of corresponding results for the Lasso (Yuan and Lin, 2007; Zhao and Yu, 2006; Zou, 2006) and the group Lasso (Bach, 2008). We do so under two sets of sampling

assumptions detailed in Section 3.2: a full i.i.d assumption and a non i.i.d assumption which is natural in the context of collaborative filtering.

As for the Lasso and the group Lasso, the necessary condition implies that such procedures do not always estimate the rank correctly; similar to the adaptive version of the Lasso and group Lasso (Zou, 2006), we design an adaptive version to achieve $n^{-1/2}$-consistency and rank consistency, with no consistency conditions. Following Zou (2006), the adaptive version is based on a unregularized least-squares estimates which is used to design appropriate reweighted matrices. Finally, in Section 6, we present a smoothing approach to convex optimization with the trace norm, while in Section 6.3, we show simulations on toy examples to illustrate the consistency results.

2. Notation

In this paper we consider various norms on vectors and matrices. On vectors x in $\mathbb{R}^d$, we always consider the Euclidean norm, that is, $\|x\| = (x^\top x)^{1/2}$. On rectangular matrices in $\mathbb{R}^{p \times q}$, however, we consider several norms, based on singular values (Stewart and Sun, 1990): the *spectral norm* $\|M\|_2$ is the largest singular value (defined as $\|M\|_2 = \sup_{x \in \mathbb{R}^q} \frac{\|Mx\|}{\|x\|}$), the *trace norm* (or *nuclear norm*) $\|M\|_*$ is the sum of singular values, and the *Frobenius norm* $\|M\|_F$ is the ℓ_2-norm of singular values—also defined as $\|M\|_F = (\mathrm{tr} M^\top M)^{1/2}$. In Appendix A and B, we review and derive relevant tools and results regarding perturbation of singular values as well as the trace norm.

Given a matrix $M \in \mathbb{R}^{p \times q}$, $\mathrm{vec}(M)$ denotes the vector in $\mathbb{R}^{pq}$ obtained by stacking its columns into a single vector; and $A \otimes B$ denotes the Kronecker product between matrices $A \in \mathbb{R}^{p_1 \times q_1}$ and $B \in \mathbb{R}^{p_2 \times q_2}$, defined as the matrix in $\mathbb{R}^{p_1 p_2 \times q_1 q_2}$, defined by blocks of sizes $p_2 \times q_2$ equal to $a_{ij}B$. We make constant use of the following identities: $(B^\top \otimes A) \mathrm{vec}(X) = \mathrm{vec}(AXB)$ and $\mathrm{vec}(uv^\top) = v \otimes u$. For more details and properties, see Golub and Loan (1996) and Magnus and Neudecker (1998). We also use the notation ΣW for $\Sigma \in \mathbb{R}^{pq \times pq}$ and $W \in \mathbb{R}^{p \times q}$ to design the matrix in $\mathbb{R}^{p \times q}$ such that $\mathrm{vec}(\Sigma W) = \Sigma \mathrm{vec}(W)$ (note the potential confusion with ΣW when Σ is a matrix with p columns).

We use the following standard asymptotic notations: a random variable Z_n is said to be of order $O_p(a_n)$ if for any $\eta > 0$, there exists $M > 0$ such that $\sup_n P(|Z_n| > M a_n) < \eta$. Moreover, Z_n is said to be of order $o_p(a_n)$ if Z_n/a_n converges to zero in probability, that is, if for any $\eta > 0$, $P(|Z_n| \geqslant \eta a_n)$ converges to zero. See Van der Vaart (1998) and Shao (2003) for further definitions and properties of asymptotics in probability.

Finally, we use the following two conventions: lowercase for vectors and uppercase for matrices, while bold fonts are reserved for population quantities.

3. Trace Norm Minimization

We consider the problem of predicting a real random variable z as a linear function of a matrix $M \in \mathbb{R}^{p \times q}$, where p and q are two fixed strictly positive integers. Throughout this paper, we assume that we are given n observations (M_i, z_i), $i = 1, \ldots, n$, and we consider the following optimization problem with the square loss:

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2n} \sum_{i=1}^{n} (z_i - \mathrm{tr} W^\top M_i)^2 + \lambda_n \|W\|_*, \tag{1}$$

where $\|W\|_*$ denotes the *trace norm* of W.

3.1 Special Cases

Regularization by the trace norm has numerous applications (see, e.g., Recht et al., 2007, for a review); in this paper, we are particularly interested in the following two situations:

Lasso and group Lasso When $x_i \in \mathbb{R}^m$, we can define $M_i = \mathrm{Diag}(x_i) \in \mathbb{R}^{m \times m}$ as the diagonal matrix with x_i on the diagonal. In this situation, the minimization of problem in Eq. (1) must lead to diagonal solutions (indeed the minimum trace norm matrix with fixed diagonal is the corresponding diagonal matrix, which is a consequence of Lemma 20 and Proposition 21) and for a diagonal matrix the trace norm is simply the ℓ_1 norm of the diagonal. Once we have derived our consistency conditions, we check in Section 4.5 that they actually lead to the known ones for the Lasso (Yuan and Lin, 2007; Zhao and Yu, 2006; Zou, 2006).

We can also see the group Lasso as a special case; indeed, if $x_{ij} \in \mathbb{R}^{d_j}$ for $j = 1, \ldots, m$, $i = 1, \ldots, n$, then we define $M_i \in \mathbb{R}^{(\Sigma_{j=1}^m d_j) \times m}$ as the block diagonal matrix (with non square blocks) with diagonal blocks x_{ji}, $j = 1, \ldots, m$. Similarly, the optimal $\hat{W}$ must share the same block-diagonal form, and its singular values are exactly the norms of each block, that is, the trace norm is indeed the sum of the norms of each group. We also get back results from Bach (2008) in Section 4.5.

Note that the Lasso and group Lasso can be seen as special cases where the singular vectors are fixed. However, the main difficulty in analyzing trace norm regularization, as well as the main reason for it use, is that singular vectors are not fixed and those can often be seen as implicit features learned by the estimation procedure (Srebro et al., 2005). In this paper we derive consistency results about the value and numbers of such features.

Collaborative filtering and low-rank completion Another natural application is collaborative filtering where two types of attributes x and y are observed and we consider bilinear forms in x and y, which can be written as a linear form in $M = xy^\top$ (thus it corresponds to situations where all matrices M_i have rank one). In this setting, the matrices M_i are not usually i.i.d. but exhibit a statistical dependence structure outlined in Section 3.2. A special case here is when then no attributes are observed and we simply wish to complete a partially observed matrix (Srebro et al., 2005; Abernethy et al., 2006). The results presented in this paper do not immediately apply because the dimension of the estimated matrix may grow with the number of observed entries and this situation is out of the scope of this paper.

Multivariate linear supervised learning When predicting multiple variables, in the context of multivariate linear regression (Yuan et al., 2007) or in the multiple category classification (Amit et al., 2007), the trace norm allows to perform feature selection.

3.2 Assumptions

We make the following assumptions on the sampling distributions of $M \in \mathbb{R}^{p \times q}$ for the problem in Eq. (1). We let denote: $\hat{\Sigma}_{mm} = \frac{1}{n} \sum_{i=1}^n \mathrm{vec}(M_i) \mathrm{vec}(M_i)^\top \in \mathbb{R}^{pq \times pq}$, and we consider the following assumptions:

(**A1**) Given M_i, $i = 1, \ldots, n$, the n values z_i are i.i.d. and there exists $\mathbf{W} \in \mathbb{R}^{p \times q}$ such that for all i, $\mathbb{E}(z_i | M_1, \ldots, M_n) = \mathrm{tr} \mathbf{W}^\top M_i$ and $\mathrm{var}(z_i | M_1, \ldots, M_n)$ is a strictly positive constant σ^2. $\mathbf{W}$ is not equal to zero and does not have full rank.

(A2) There exists an *invertible* matrix $\Sigma_{mm} \in \mathbb{R}^{pq \times pq}$ such that $\mathbb{E}\|\hat{\Sigma}_{mm} - \Sigma_{mm}\|_F^2 = O(\zeta_n^2)$ for a certain sequence ζ_n that tends to zero.

(A3) The random variable $n^{-1/2}\sum_{i=1}^n \varepsilon_i \operatorname{vec}(M_i)$ is converging in distribution to a normal distribution with mean zero and covariance matrix $\sigma^2 \Sigma_{mm}$.

Assumption (**A1**) states that given the input matrices M_i, $i = 1,\ldots,n$ we have a linear prediction model, where the loading matrix $\mathbf{W}$ is non trivial and rank-deficient, the goal being to estimate this rank (as well as the matrix itself). We let denote $\mathbf{W} = \mathbf{U}\operatorname{Diag}(\mathbf{s})\mathbf{V}^\top$ its singular value decomposition, with $\mathbf{U} \in \mathbb{R}^{p \times \mathbf{r}}$, $\mathbf{V} \in \mathbb{R}^{q \times \mathbf{r}}$, and $\mathbf{r} \in (0, \min\{p,q\})$ denotes the rank of $\mathbf{W}$. We also let denote $\mathbf{U}_\perp \in \mathbb{R}^{p \times (p-\mathbf{r})}$ and $\mathbf{V}_\perp \in \mathbb{R}^{q \times (q-\mathbf{r})}$ any orthogonal complements of $\mathbf{U}$ and $\mathbf{V}$.

We let denote $\varepsilon_i = z_i - \operatorname{tr}\mathbf{W}^\top M_i$ and $\hat{\Sigma}_{Mz} = \frac{1}{n}\sum_{i=1}^n z_i M_i \in \mathbb{R}^{p \times q}$, $\hat{\Sigma}_{M\varepsilon} = \frac{1}{n}\sum_{i=1}^n \varepsilon_i M_i = \hat{\Sigma}_{Mz} - \hat{\Sigma}_{mm}\mathbf{W} \in \mathbb{R}^{p \times q}$. We may then rewrite Eq. (1) as

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2}\operatorname{vec}(W)^\top \hat{\Sigma}_{mm}\operatorname{vec}(W) - \operatorname{tr}W^\top \hat{\Sigma}_{Mz} + \lambda_n\|W\|_*,$$

or, equivalently,

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2}\operatorname{vec}(W - \mathbf{W})^\top \hat{\Sigma}_{mm}\operatorname{vec}(W - \mathbf{W}) - \operatorname{tr}W^\top \hat{\Sigma}_{M\varepsilon} + \lambda_n\|W\|_*.$$

The sampling assumptions (**A2**) and (**A3**) may seem restrictive, but they are satisfied in the following two natural situations. The first situation corresponds to a classical full i.i.d problem, where the pairs (z_i, M_i) are sampled i.i.d:

Lemma 1 *Assume (**A1**). If the matrices M_i are sampled i.i.d., z and M have finite fourth order moments, and $\mathbb{E}\{\operatorname{vec}(M)\operatorname{vec}(M)^\top\}$ is invertible, then (**A2**) and (**A3**) are satisfied with $\zeta_n = n^{-1/2}$.*

Note the further refinement when for each i, $M_i = x_i y_i^\top$ and x_i and y_i are independent, which implies that Σ_{mm} is factorized as a Kronecker product, of the form $\Sigma_{yy} \otimes \Sigma_{xx}$ where Σ_{xx} and Σ_{yy} are the (invertible) second order moment matrices of x and y.

The second situation corresponds to a collaborative filtering situation where two types of attributes are observed, for example, x and y, and for every pair (x,y) we wish to predict z as a bilinear form in x and y: we first sample n_x values for x, and n_y values for y, and we select uniformly at random a subset of $n \leqslant n_x n_y$ observations from the $n_x n_y$ possible pairs. The following lemma, proved in Appendix C.1, shows that this set-up satisfies our assumptions:

Lemma 2 *Assume (**A1**). Assume moreover that n_x values $\tilde{x}_1,\ldots,\tilde{x}_{n_x}$ are sampled i.i.d and n_y values $\tilde{y}_1,\ldots,\tilde{y}_{n_y}$ are also sampled i.i.d. from distributions with finite fourth order moments and invertible second order moment matrices Σ_{xx} and Σ_{yy}. Assume also that a random subset of size n of pairs (i_k, j_k) in $\{1,\ldots,n_x\} \times \{1,\ldots,n_y\}$ is sampled uniformly, then if n_x, n_y and n tend to infinity, then (**A2**) and (**A3**) are satisfied with $\Sigma_{mm} = \Sigma_{yy} \otimes \Sigma_{xx}$ and $\zeta_n = n^{-1/2} + n_x^{-1/2} + n_y^{-1/2}$.*

3.3 Optimality Conditions

From the expression of the subdifferential of the trace norm in Proposition 21 (Appendix B), we can identify the optimality condition for problem in Eq. (1), that we will constantly use in the paper:

Proposition 3 *The matrix W with singular value decomposition $W = U \operatorname{Diag}(s) V^\top$ (with strictly positive singular values s) is optimal for the problem in Eq. (1) if and only if*

$$\hat{\Sigma}_{mm} W - \hat{\Sigma}_{Mz} + \lambda_n U V^\top + N = 0,$$

with $U^\top N = 0$, $NV = 0$ and $\|N\|_2 \leqslant \lambda_n$.

This implies notably that W and $\hat{\Sigma}_{mm} W - \hat{\Sigma}_{Mz}$ have *simultaneous* singular value decompositions, and the largest singular values are less than λ_n, and exactly equal to λ_n for the corresponding strictly positive singular values of W. Note that when all matrices are diagonal (the Lasso case), we obtain the usual optimality conditions (see also Recht et al., 2007, for further discussions).

4. Consistency Results

We consider two types of consistency; first, the *regular consistency*, that is, we want the probability $\mathbb{P}(\|\hat{W} - \mathbf{W}\| \geqslant \varepsilon)$ to tend to zero as n tends to infinity, for all $\varepsilon > 0$. We also consider the *rank consistency*, that is, we want that $\mathbb{P}(\operatorname{rank}(\hat{W}) \neq \operatorname{rank}(\mathbf{W}))$ tends to zero as n tends to infinity. Following the similar properties for the Lasso, the consistency depends on the decay of the regularization parameter. Essentially, we obtain the following results:

a) if λ_n does not tend to zero, then the trace norm estimate $\hat{W}$ is not consistent;

b) if λ_n tends to zero faster than $n^{-1/2}$, then the estimate is consistent and its error is $O_p(n^{-1/2})$ while it is not rank-consistent with probability tending to one (see Section 4.1);

c) if λ_n tends to zero exactly at rate $n^{-1/2}$, then the estimator is consistent with error $O_p(n^{-1/2})$ but the probability of estimating the correct rank is converging to a limit in $(0,1)$ (see Section 4.2);

d) if λ_n tends to zero more slowly than $n^{-1/2}$, then the estimate is consistent with error $O_p(\lambda_n)$ and its rank consistency depends on specific *consistency conditions* detailed in Section 4.3.

The following sections will look at each of these cases, and state precise theorems. We then consider some special cases, that is, factored second-order moments and implications for the special cases of the Lasso and group Lasso.

The first proposition (proved in Appendix C.2) considers the case where the regularization parameter λ_n is converging to a certain limit λ_0. When this limit is zero, we obtain regular consistency (Corollary 5 below), while if $\lambda_0 > 0$, then $\hat{W}$ tends in probability to a limit which is always different from $\mathbf{W}$:

Proposition 4 *Assume (**A1-3**). Let $\hat{W}$ be a global minimizer of Eq. (1). If λ_n tends to a limit $\lambda_0 \geqslant 0$, then $\hat{W}$ converges in probability to the unique global minimizer of*

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2} \operatorname{vec}(W - \mathbf{W})^\top \Sigma_{mm} \operatorname{vec}(W - \mathbf{W}) + \lambda_0 \|W\|_*.$$

Corollary 5 *Assume (**A1-3**). Let $\hat{W}$ be a global minimizer of Eq. (1). If λ_n tends to zero, then $\hat{W}$ converges in probability to $\mathbf{W}$.*

We now consider finer results when λ_n tends to zero at certain rates, slower or faster than $n^{-1/2}$, or exactly at rate $n^{-1/2}$.

4.1 Fast Decay of Regularization Parameter

The following proposition—which is a consequence of standard results in M-estimation (Shao, 2003; Van der Vaart, 1998)—considers the case where $n^{1/2}\lambda_n$ is tending to zero, where we obtain that $\hat{W}$ is asymptotically normal with mean $\mathbf{W}$ and covariance matrix $n^{-1}\sigma^2\Sigma_{mm}^{-1}$, that is, for fast decays, the first order expansion is the same as the one with no regularization parameter:

Proposition 6 *Assume (***A1-3***). Let $\hat{W}$ be a global minimizer of Eq. (1). If $n^{1/2}\lambda_n$ tends to zero, $n^{1/2}(\hat{W} - \mathbf{W})$ is asymptotically normal with mean $\mathbf{W}$ and covariance matrix $\sigma^2\Sigma_{mm}^{-1}$.*

We now consider the corresponding rank consistency results, when λ_n goes to zero faster than $n^{-1/2}$. The following proposition (proved in Appendix C.3) states that for such regularization parameter, the solution has rank strictly greater than $\mathbf{r}$ with probability tending to one and can thus not be rank consistent:

Proposition 7 *Assume (***A1-3***). If $n^{1/2}\lambda_n$ tends to zero, then $\mathbb{P}(\text{rank}(\hat{W}) > \text{rank}(\mathbf{W}))$ tends to one.*

4.2 $n^{-1/2}$-decay of the Regularization Parameter

We first consider regular consistency through the following proposition (proved in Appendix C.4), then rank consistency (proposition proved in Appendix C.5):

Proposition 8 *Assume (***A1-3***). Let $\hat{W}$ be a global minimizer of Eq. (1). If $n^{1/2}\lambda_n$ tends to a limit $\lambda_0 > 0$, then $n^{1/2}(\hat{W} - \mathbf{W})$ converges in distribution to the unique global minimizer of*

$$\min_{\Delta \in \mathbb{R}^{p \times q}} \frac{1}{2} \text{vec}(\Delta)^\top \Sigma_{mm} \text{vec}(\Delta) - \text{tr}\Delta^\top A + \lambda_0 \left[\text{tr}\mathbf{U}^\top \Delta \mathbf{V} + \|\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp\|_* \right],$$

where $\text{vec}(A) \in \mathbb{R}^{pq}$ is normally distributed with mean zero and covariance matrix $\sigma^2 \Sigma_{mm}$.

Proposition 9 *Assume (***A1-3***). If $n^{1/2}\lambda_n$ tends to a limit $\lambda_0 > 0$, then the probability that the rank of $\hat{W}$ is different from the rank of $\mathbf{W}$ is converging to $\mathbb{P}(\|\Lambda - \lambda_0^{-1}\Theta\|_2 \leqslant 1) \in (0,1)$ where $\Lambda \in \mathbb{R}^{(p-\mathbf{r}) \times (q-\mathbf{r})}$ is defined in Eq. (3) (Section 4.3) and $\Theta \in \mathbb{R}^{(p-\mathbf{r}) \times (q-\mathbf{r})}$ has a normal distribution with mean zero and covariance matrix*

$$\sigma^2 \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1} (\mathbf{V}_\perp \otimes \mathbf{U}_\perp) \right)^{-1}.$$

The previous proposition ensures that the estimate $\hat{W}$ cannot be rank consistent with this decay of the regularization parameter. Note that when we take λ_0 small (i.e., we get closer to fast decays), the probability $\mathbb{P}(\|\Lambda - \lambda_0^{-1}\Theta\|_2 \leqslant 1)$ tends to zero, while when we take λ_0 large (i.e., we get closer to slow decays), the same probability tends to zero or one depending on the sign of $\|\Lambda\|_2 - 1$. This heuristic argument is made more precise in the following section.

4.3 Slow Decay of Regularization Parameter

When λ_n tends to zero more slowly than $n^{-1/2}$, the first order expansion is deterministic, as the following proposition shows (proof in Appendix C.6):

Proposition 10 *Assume (**A1-3**). Let $\hat{W}$ be a global minimizer of Eq. (1). If $n^{1/2}\lambda_n$ tends to $+\infty$ and λ_n tends to zero, then $\lambda_n^{-1}(\hat{W} - \mathbf{W})$ converges in probability to the unique global minimizer Δ of*

$$\min_{\Delta \in \mathbb{R}^{p \times q}} \frac{1}{2} \text{vec}(\Delta)^\top \Sigma_{mm} \text{vec}(\Delta) + \text{tr}\mathbf{U}^\top \Delta \mathbf{V} + \|\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp\|_*. \tag{2}$$

Moreover, we have $\hat{W} = \mathbf{W} + \lambda_n \Delta + O_p(\lambda_n + \zeta_n + \lambda_n^{-1} n^{-1/2})$.

The last proposition gives a first order expansion of $\hat{W}$ around $\mathbf{W}$. From Proposition 18 (Appendix B), we obtain immediately that if $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp$ is different from zero, then the rank of $\hat{W}$ is ultimately strictly larger than $\mathbf{r}$. The condition $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp = 0$ is thus necessary for rank consistency when $\lambda_n n^{1/2}$ tends to infinity while λ_n tends to zero. The next lemma (proved in Appendix 11), gives a necessary and sufficient condition for $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp = 0$.

Lemma 11 *Assume Σ_{mm} is invertible, and $\mathbf{W} = \mathbf{U}\text{Diag}(\mathbf{s})\mathbf{V}^\top$ is the singular value decomposition of $\mathbf{W}$. Then the unique global minimizer of*

$$\text{vec}(\Delta)^\top \Sigma_{mm} \text{vec}(\Delta) + \text{tr}\mathbf{U}^\top \Delta \mathbf{V} + \|\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp\|_*$$

satisfies $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp = 0$ if and only if

$$\left\| \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1} (\mathbf{V}_\perp \otimes \mathbf{U}_\perp) \right)^{-1} \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1} (\mathbf{V} \otimes \mathbf{U}) \text{vec}(\mathbf{I}) \right) \right\|_2 \leqslant 1.$$

This leads to consider the matrix $\Lambda \in \mathbb{R}^{(p-\mathbf{r}) \times (q-\mathbf{r})}$ defined as

$$\text{vec}(\Lambda) = \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1} (\mathbf{V}_\perp \otimes \mathbf{U}_\perp) \right)^{-1} \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1} (\mathbf{V} \otimes \mathbf{U}) \text{vec}(\mathbf{I}) \right), \tag{3}$$

and the two weak and strict *consistency conditions*:

$$\|\Lambda\|_2 \leqslant 1, \tag{4}$$

$$\|\Lambda\|_2 < 1. \tag{5}$$

Note that if Σ_{mm} is proportional to identity, they are always satisfied because then $\Lambda = 0$. We can now prove that the condition in Eq. (5) is sufficient for rank consistency when $n^{1/2}\lambda_n$ tends to infinity, while the condition Eq. (4) is necessary for the existence of a sequence λ_n such that the estimate is both consistent and rank consistent (which is a stronger result than restricting λ_n to be tending to zero slower than $n^{-1/2}$). The following two theorems are proved in Appendix C.8 and C.9:

Theorem 12 *Assume (**A1-3**). Let $\hat{W}$ be a global minimizer of Eq. (1). If the condition in Eq. (5) is satisfied, and if $n^{1/2}\lambda_n$ tends to $+\infty$ and λ_n tends to zero, then the estimate $\hat{W}$ is consistent and rank-consistent.*

Theorem 13 *Assume (**A1-3**). Let $\hat{W}$ be a global minimizer of Eq. (1). If the estimate $\hat{W}$ is consistent and rank-consistent, then the condition in Eq. (4) is satisfied.*

As opposed to the Lasso, where Eq. (4) is a necessary and sufficient condition for rank consistency (Yuan and Lin, 2007), this is not even true in general for the group Lasso (Bach, 2008). Looking at the limiting case $\|\Lambda\|_2 = 1$ would similarly lead to additional but more complex sufficient and necessary conditions, and is left out for future research.

Moreover, it may seem surprising that even when the sufficient condition Eq. (5) is fulfilled, that the first order expansion of $\hat{W}$, that is, $\hat{W} = \mathbf{W} + \lambda_n \Delta + o_p(\lambda_n)$ is such that $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp = 0$, but nothing is said about $\mathbf{U}_\perp^\top \Delta \mathbf{V}$ and $\mathbf{U}^\top \Delta \mathbf{V}_\perp$, which are not equal to zero in general. This is due to the fact that the first $\mathbf{r}$ singular vectors U and V of $\mathbf{W} + \lambda_n \Delta$ are not fixed; indeed, the $\mathbf{r}$ first singular vectors (i.e., the implicit features) do rotate but with no contribution on $\mathbf{U}_\perp \mathbf{V}_\perp^\top$. This is to be contrasted with the adaptive version where asymptotically the first order expansion has constant singular vectors (see Section 5).

Finally, in this paper, we have only proved whether the probability of correct rank selection tends to zero or one. Proposition 9 suggests that when $\lambda_n n^{1/2}$ tends to infinity slowly, then this probability is close to $\mathbb{P}(\|\Lambda - \lambda_n^{-1} n^{1/2} \Theta\|_2 \leqslant 1)$, where Θ has a normal distribution with known covariance matrix, which converges to one exponentially fast when $\|\Lambda\|_2 < 1$. We are currently investigating additional assumptions under which such results are true and thus estimate the convergence rates of the probability of good rank selection as done by Zhao and Yu (2006) for the Lasso.

4.4 Factored Second Order Moment

Note that in the situation where n_x points in $\mathbb{R}^p$ and n_y points in $\mathbb{R}^q$ are sampled i.i.d and a random subset of n points in selected, then, we can refine the condition as follows (because $\Sigma_{mm} = \Sigma_{yy} \otimes \Sigma_{xx}$):

$$\Lambda = (\mathbf{U}_\perp^\top \Sigma_{xx}^{-1} \mathbf{U}_\perp)^{-1} \mathbf{U}_\perp^\top \Sigma_{xx}^{-1} \mathbf{U} \mathbf{V}^\top \Sigma_{yy}^{-1} \mathbf{V}_\perp (\mathbf{V}_\perp^\top \Sigma_{yy}^{-1} \mathbf{V}_\perp)^{-1},$$

which is equal to (by the expression of inverses of partitioned matrices):

$$\Lambda = (\mathbf{U}_\perp^\top \Sigma_{xx} \mathbf{U})(\mathbf{U}^\top \Sigma_{xx} \mathbf{U})^{-1} (\mathbf{V}^\top \Sigma_{yy} \mathbf{V})^{-1} (\mathbf{V}^\top \Sigma_{yy} \mathbf{V}_\perp).$$

This also happens when $M_i = x_i y_i^\top$ and x_i and y_i independent for all i.

4.5 Corollaries for the Lasso and Group Lasso

For the Lasso or the group Lasso, all proposed results in Section 4.3 should hold with the additional conditions that W and Δ are diagonal (block-diagonal for the group Lasso). In this situation, the singular values of the diagonal matrix $W = \mathrm{Diag}(w)$ are the norms of the diagonal blocks, while the left singular vectors are equal to the normalized versions of the block (the signs for the Lasso). However, the results developed in Section 4.3 do not immediately apply since the assumptions regarding the invertibility of the second order moment matrix is not satisfied. For those problems, all matrices M that are ever considered belong to a strict subspace of $\mathbb{R}^{p \times q}$ and we need to satisfy invertibility on that subspace.

More precisely, we assume that all matrices M are such that $\mathrm{vec}(M) = Hx$ where H is a given *design matrix* in $\mathbb{R}^{pq \times s}$ where s is the number of implicit parameter and $x \in \mathbb{R}^s$. If we replace the invertibility of Σ_{mm} by the invertibility of $H^\top \Sigma_{mm} H$, then all results presented in Section 4.3 are

valid, in particular, the matrix Λ may be written as

$$\text{vec}(\Lambda) = \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top H (H^\top \Sigma_{mm} H)^{-1} H^\top (\mathbf{V}_\perp \otimes \mathbf{U}_\perp) \right)^\dagger$$
$$\times \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top H (H^\top \Sigma_{mm} H)^{-1} H^\top (\mathbf{V} \otimes \mathbf{U}) \text{vec}(\mathbf{I}) \right), \quad (6)$$

where $A^\dagger$ denotes the pseudo-inverse of A (Golub and Loan, 1996).

We now apply Eq. (6) to the case of the group Lasso (which includes the Lasso as a special case). In this situation, we have $M = \text{Diag}(x_1, \ldots, x_m)$ and each $x_j \in \mathbb{R}^{d_j}$, $j = 1, \ldots, m$; we consider w as being defined by blocks $w_1, \ldots, w_m$, where each $w_j \in \mathbb{R}^{d_j}$. The design matrix H is such that $Hw = \text{vec}(\text{Diag}(w))$ and the matrix $H^\top \Sigma_{mm} H$ is exactly equal to the joint covariance matrix Σ_{xx} of $x = (x_1, \ldots, x_m)$. Without loss of generality, we assume that the generating sparsity pattern corresponds to the first $\mathbf{r}$ blocks. We can then compute the singular value decomposition in closed form as $\mathbf{U} = \begin{pmatrix} (\text{Diag}(\mathbf{w}_i/\|\mathbf{w}_i\|)_{i \leqslant \mathbf{r}}) \\ 0 \end{pmatrix}$, $\mathbf{V} = \begin{pmatrix} \mathbf{I} \\ 0 \end{pmatrix}$ and $\mathbf{s} = (\|\mathbf{w}_j\|)_{j \leqslant \mathbf{r}}$. If we let denote, for each j, $\mathbf{O}_j$ a basis of the subspace orthogonal to $\mathbf{w}_j$, we have: $\mathbf{U}_\perp = \begin{pmatrix} \text{Diag}(\mathbf{O}_i)_{i \leqslant \mathbf{r}} & 0 \\ 0 & \mathbf{I} \end{pmatrix}$ and $\mathbf{V}_\perp = \begin{pmatrix} 0 \\ \mathbf{I} \end{pmatrix}$. We can put these singular vectors into Eq. (6) and get $(H^\top \Sigma_{mm} H)^{-1} H^\top (\mathbf{V} \otimes \mathbf{U}) \text{vec}(\mathbf{I}) = (\Sigma_{xx}^{-1})_{\mathbf{J} \cup \mathbf{J}^c, \mathbf{J}} \eta_{\mathbf{J}}$, where $\mathbf{J} = \{1, \ldots, \mathbf{r}\}$ and $\eta_{\mathbf{J}}$ is the vector of normalized $\mathbf{w}_j$, $j \in \mathbf{J}$. Thus, for the group Lasso, we finally obtain:

$$\begin{aligned} \|\Lambda\|_2 &= \left\| \text{Diag} \left[((\Sigma_{xx}^{-1})_{\mathbf{J}^c \mathbf{J}^c})^{-1} (\Sigma_{xx}^{-1})_{\mathbf{J}^c, \mathbf{J}} \eta_{\mathbf{J}} \right] \right\|_2 \\ &= \left\| \text{Diag} \left[(\Sigma_{xx})_{\mathbf{J}^c \mathbf{J}} (\Sigma_{xx})_{\mathbf{J}, \mathbf{J}}^{-1} \eta_{\mathbf{J}} \right] \right\|_2 \text{ by the partitioned matrices inversion lemma,} \\ &= \max_{i \in \mathbf{J}^c} \left\| \Sigma_{x_i x_{\mathbf{J}}} \Sigma_{x_{\mathbf{J}} x_{\mathbf{J}}}^{-1} \eta_{\mathbf{J}} \right\|. \end{aligned}$$

The condition on the invertibility of $H^\top \Sigma_{mm} H$ is exactly the invertibility of the full joint covariance matrix of $x = (x_1, \ldots, x_m)$ and is a standard assumption for the Lasso or the group Lasso (Yuan and Lin, 2007; Zhao and Yu, 2006; Zou, 2006; Bach, 2008). Moreover the condition $\|\Lambda\|_2 \leqslant 1$ is exactly the one for the group Lasso (Bach, 2008), where the pattern consistency is replaced by the consistency for the number of non zero groups.

Note that we only obtain a result in terms of numbers of selected groups of variables and not in terms of the identities of the groups themselves. However, because of regular consistency, we know that at least the $\mathbf{r}$ true groups will be selected, and then correct model size is asymptotically equivalent to the correct groups being selected.

5. Adaptive Version

We can follow the adaptive version of the Lasso to provide a consistent algorithm with no consistency conditions such as Eq. (4) or Eq. (5). More precisely, we consider the least-square estimate $\text{vec}(\hat{W}_{LS}) = \hat{\Sigma}_{mm}^{-1} \text{vec}(\hat{\Sigma}_{Mz})$. We have the following well known result for least-square regression:

Lemma 14 *Assume (A1-3). Then $n^{1/2}(\hat{\Sigma}_{mm}^{-1} \text{vec}(\hat{\Sigma}_{Mz}) - \text{vec}(\mathbf{W}))$ is converging in distribution to a normal distribution with zero mean and covariance matrix $\sigma^2 \Sigma_{mm}^{-1}$.*

We consider the singular value decomposition of $\hat{W}_{LS} = U_{LS} \text{Diag}(s_{LS}) V_{LS}^\top$, where $s_{LS} \geqslant 0$. With probability tending to one, $\min\{p, q\}$ singular values are strictly positive (i.e., the rank of $\hat{W}_{LS}$ is

full). We consider the *full* decomposition where U_{LS} and V_{LS} are orthogonal *square* matrices and the matrix $\mathrm{Diag}(s_{LS})$ is rectangular. We complete the singular values $s_{LS} \in \mathbb{R}^{\min\{p,q\}}$ by $n^{-1/2}$ to reach dimensions p and q (we keep the same notation for both dimensions for simplicity).

For $\gamma \in (0,1]$, we let denote

$$A = U_{LS}\,\mathrm{Diag}(s_{LS})^{-\gamma}U_{LS}^{\top} \in \mathbb{R}^{p \times p} \text{ and } B = V_{LS}\,\mathrm{Diag}(s_{LS})^{-\gamma}V_{LS}^{\top} \in \mathbb{R}^{q \times q},$$

two positive definite symmetric matrices, and, following the adaptive Lasso of Zou (2006), we consider replacing $\|W\|_*$ by $\|AWB\|_*$—note that in the Lasso special case, this exactly corresponds to the adaptive Lasso of Zou (2006). We obtain the following consistency theorem (proved in Appendix C.10):

Theorem 15 *Assume (**A1-3**). If $\gamma \in (0,1]$, $n^{1/2}\lambda_n$ tends to 0 and $\lambda_n n^{1/2+\gamma/2}$ tends to infinity, then any global minimizer $\hat{W}_A$ of*

$$\frac{1}{2n}\sum_{i=1}^{n}(z_i - \mathrm{tr}W^{\top}M_i)^2 + \lambda_n\|AWB\|_*$$

is consistent and rank consistent. Moreover, $n^{1/2}\,\mathrm{vec}(\hat{W}_A - \mathbf{W})$ is converging in distribution to a normal distribution with mean zero and covariance matrix

$$\sigma^2(\mathbf{V}\otimes\mathbf{U})\left[(\mathbf{V}\otimes\mathbf{U})^{\top}\Sigma_{mm}(\mathbf{V}\otimes\mathbf{U})\right]^{-1}(\mathbf{V}\otimes\mathbf{U})^{\top}.$$

Note the restriction $\gamma \leqslant 1$ which is due to the fact that the least-square estimate $\hat{W}_{LS}$ only estimates the singular subspaces at rate $O_p(n^{-1/2})$. In Section 6.3, we illustrate the previous theorem on synthetic examples. In particular, we exhibit some singular behavior for the limiting case $\gamma = 1$.

6. Algorithms and Simulations

In this section we provide a simple algorithm to solve problems of the form

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2}\mathrm{vec}(W)^{\top}\Sigma\,\mathrm{vec}(W) - \mathrm{tr}W^{\top}Q + \lambda\|W\|_*, \tag{7}$$

where $\Sigma \in \mathbb{R}^{pq \times pq}$ is a positive definite matrix (note that we do not restrict Σ to be of the form $\Sigma = A \otimes B$ where A and B are positive semidefinite matrices of size $p \times p$ and $q \times q$). We assume that $\mathrm{vec}(Q)$ is in the column space of Σ, so that the optimization problem is bounded from below (and thus the dual is feasible). In our setting, we have $\Sigma = \hat{\Sigma}_{mm}$ and $Q = \hat{\Sigma}_{Mz}$.

We focus on problems where p and q are not too large so that we can apply Newton's method to obtain convergence up to machine precision, which is required for the fine analysis of rank consistency in Section 6.3. For more efficient algorithms with larger p and q, see Srebro et al. (2005); Rennie and Srebro (2005) and Abernethy et al. (2006); Lu et al. (2008).

Because the dual norm of the trace norm is the spectral norm (see Appendix B), the dual is easily obtained as

$$\max_{V \in \mathbb{R}^{p \times q}, \|V\|_2 \leqslant 1} -\frac{1}{2}\mathrm{vec}(Q - \lambda V)^{\top}\Sigma^{-1}\mathrm{vec}(Q - \lambda V). \tag{8}$$

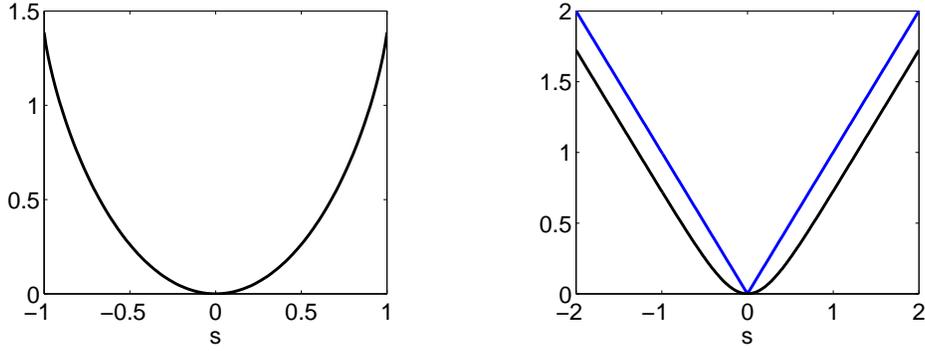Figure 1: Spectral barrier functions: (left) primal function $b(s)$ and (right) dual functions $b^*(s)$.

Indeed, we have:

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2} \text{vec}(W)^\top \Sigma \text{vec}(W) - \text{tr} W^\top Q + \lambda \|W\|_*$$

$$= \min_{W \in \mathbb{R}^{p \times q}} \max_{V \in \mathbb{R}^{p \times q}, \|V\|_2 \leqslant 1} \frac{1}{2} \text{vec}(W)^\top \Sigma \text{vec}(W) - \text{tr} W^\top Q + \lambda \text{tr} V^\top W$$

$$= \max_{V \in \mathbb{R}^{p \times q}, \|V\|_2 \leqslant 1} \min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2} \text{vec}(W)^\top \Sigma \text{vec}(W) - \text{tr} W^\top Q + \lambda \text{tr} V^\top W$$

$$= \max_{V \in \mathbb{R}^{p \times q}, \|V\|_2 \leqslant 1} -\frac{1}{2} \text{vec}(Q - \lambda V)^\top \Sigma^{-1} \text{vec}(Q - \lambda V),$$

where strong duality holds because both the primal and dual problems are convex and strictly feasible (Boyd and Vandenberghe, 2003).

6.1 Smoothing

The problem in Eq. (7) is convex but non differentiable; in this paper we consider adding a strictly convex function to its dual in Eq. (8) in order to make it differentiable, while controlling the increase of duality gap yielded by the added function (Bonnans et al., 2003).

We thus consider the following smoothing of the trace norm, namely we define

$$F_\varepsilon(W) = \max_{V \in \mathbb{R}^{p \times q}, \|V\|_2 \leqslant 1} \text{tr} V^\top W - \varepsilon B(V),$$

where $B(V)$ is a spectral function (i.e., that depends only on singular values of V, equal to $B(V) = \sum_{i=1}^{\min\{p,q\}} b(s_i(V))$ where $b(s) = (1+s)\log(1+s) + (1-s)\log(1-s)$ if $|s| \leqslant 1$ and $+\infty$ otherwise ($s_i(V)$ denotes the i-th largest singular values of V). This function F_ε may be computed in closed form as:

$$F_\varepsilon(W) = \sum_{i=1}^{\min\{p,q\}} b^*(s_i(W)),$$

where $b^*(s) = \varepsilon \log(1 + e^{v/\varepsilon}) + \varepsilon \log(1 + e^{-v/\varepsilon}) - 2\varepsilon \log 2$. These functions are plotted in Figure 1; note that $|b^*(s) - |s||$ is uniformly bounded by $2 \log 2$.

We finally get the following pairs of primal/dual optimization problems:

$$\min_{W \in \mathbb{R}^{p \times q}} \frac{1}{2} \mathrm{vec}(W)^\top \Sigma \, \mathrm{vec}(W) - \mathrm{tr} W^\top Q + \lambda F_{\varepsilon/\lambda}(W),$$

$$\max_{V \in \mathbb{R}^{p \times q}, \|V\|_2 \leqslant 1} -\frac{1}{2} \mathrm{vec}(Q - \lambda V)^\top \Sigma^{-1} \mathrm{vec}(Q - \lambda V) - \varepsilon B(V).$$

We can now optimize directly in the primal formulation which is infinitely differentiable, using Newton's method. Note that the stopping criterion should be an $\varepsilon \times \min\{p,q\}$ duality gap, as the controlled smoothing also leads to a small additional gap on the solution of the original non smoothed problem. More precisely, a duality gap of $\varepsilon \times \min\{p,q\}$ on the smoothed problem, leads to a gap of at most $(1 + 2\log 2)\varepsilon \times \min\{p,q\}$ for the original problem.

6.2 Implementation Details

In this section, we provide details about the implementation of the estimation algorithm presented earlier.

Derivatives of spectral functions Note that derivatives of spectral functions of the form $B(W) = \sum_{i=1}^{\min\{p,q\}} b(s_i(W))$, where b is an even twice differentiable function such that $b(0) = b'(0) = 0$, are easily calculated as follows; Let $U \mathrm{Diag}(s) V^\top$ be the singular value decomposition of W. We then have the following Taylor expansion (Lewis and Sendov, 2002):

$$B(W + \Delta) = B(W) + \mathrm{tr} \Delta^\top U \mathrm{Diag}(b'(s_i)) V^\top + \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{b'(s_i) - b'(s_j)}{s_i - s_j} (u_i^\top \Delta v_j)^2,$$

where the vector of singular values is completed by zeros, and $\frac{b'(s_i) - b'(s_j)}{s_i - s_j}$ is defined as $b''(s_i)$ when $s_i = s_j$.

Choice of ε and computational complexity Following the common practice in barrier methods we decrease the parameter geometrically after each iteration of Newton's method (Boyd and Vandenberghe, 2003). Each of these Newton iterations has complexity $O(p^3 q^3)$. Empirically, the number of iterations does not exceed a few hundreds for solving one problem up to machine precision. We are currently investigating theoretical bounds on the number of iterations through self concordance theory (Boyd and Vandenberghe, 2003).

Start and end of the path In order to avoid to consider useless values of the regularization parameter and thus use a well adapted grid for trying several λ's, we can consider a specific interval for λ. When λ is large, the solution is exactly zero, while when λ is small, the solution tends to $\mathrm{vec}(W) = \Sigma^{-1} \mathrm{vec}(Q)$.

More precisely, if λ is larger than $\|Q\|_2$, then the solution is exactly zero (because in this situation 0 is in the subdifferential). On the other side, we consider for which λ, $\Sigma^{-1} \mathrm{vec}(Q)$ leads to a duality gap which is less than $\varepsilon \, \mathrm{vec}(Q)^\top \Sigma^{-1} \mathrm{vec}(Q)$, where ε is small. A looser condition is to take $V = 0$, and the condition becomes $\lambda \|\Sigma^{-1} \mathrm{vec}(Q)\|_* \leqslant \varepsilon \, \mathrm{vec}(Q)^\top \Sigma^{-1} \mathrm{vec}(Q)$. Note that this is in the correct order (i.e., lower bound smaller than upper bound), because

$$\mathrm{vec}(Q)^\top \Sigma^{-1} \mathrm{vec}(Q) = \langle \mathrm{vec}(Q), \Sigma^{-1} \mathrm{vec}(Q) \rangle \leqslant \|\Sigma^{-1} \mathrm{vec}(Q)\|_* \|\mathrm{vec}(Q)\|_2.$$
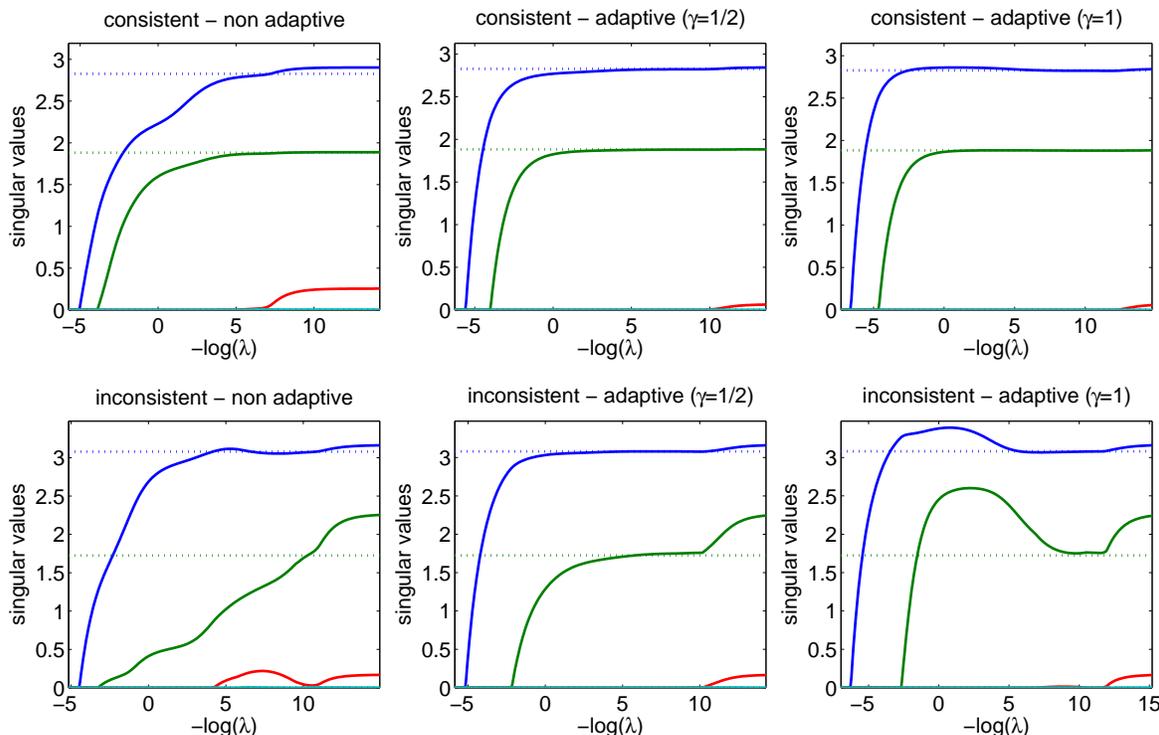
Figure 2: Examples of paths of singular values for $\|\Lambda\|_2 = 0.49 < 1$ (consistent, top) and $\|\Lambda\|_2 = 4.78 > 1$ (inconsistent, bottom) rank selection: regular trace norm penalization (left) and adaptive penalization with $\gamma = 1/2$ (center) and $\gamma = 1$ (right). Estimated singular values are plotted in plain, while population singular values are dotted.

This allows to design a good interval for searching for a good value of λ or for computing the regularization path by uniform grid sampling (in log scale), or for numerical path following with predictor-corrector methods such as used by Bach et al. (2004).

6.3 Simulations

In this section, we perform simulations on toy examples to illustrate our consistency results. We generate random i.i.d. data $\tilde{X}$ and $\tilde{Y}$ with Gaussian distributions and we select a low rank matrix $\mathbf{W}$ at random and generate $Z = \text{diag}(\tilde{X}^\top \mathbf{W} \tilde{Y}) + \varepsilon$ where ε has i.i.d components with normal distributions with zero mean and known variance. In this section, we always use $\mathbf{r} = 2$, $p = q = 4$, while we consider several numbers of samples n, and several distributions for which the consistency conditions Eq. (4) and Eq. (5) may or may not be satisfied.[1]

In Figure 2, we plot regularization paths for $n = 10^3$, by showing the singular values of $\hat{W}$ compared to the singular values of $\mathbf{W}$, in two particular situations (Eq. (4) and Eq. (5) satisfied and not satisfied), for the regular trace norm regularization and the adaptive versions, with $\gamma = 1/2$ and $\gamma = 1$. Note that in the consistent case (top), the singular values and their cardinalities are well jointly estimated, both for the non adaptive version (as predicted by Theorem 12) and the adaptive

---

1. Simulations may be reproduced with MATLAB code available from `http://www.di.ens.fr/~fbach/tracenorm/`.

versions (Theorem 15), while the range of correct rank selection increases compared to the adaptive versions. However in the inconsistent case, the non adaptive regularizations scheme (bottom left) cannot achieve regular consistency together with rank consistency (Theorem 13), while the adaptive schemes can. Note the particular behavior of the limiting case $\gamma = 1$, which still achieves both consistencies but with a singular behavior for large λ.

In Figure 3, we select the distribution used for the rank-consistent case of Figure 2, and compute the paths from 200 replications for $n = 10^2$, 10^3, 10^3 and 10^5. For each λ, we plot the proportion of estimates with correct rank on the left plots (i.e., we get an estimation of $\mathbb{P}(\text{rank}(\hat{W}) = \text{rank}(\mathbf{W}))$, while we plot the logarithm of the average root mean squared estimation error $\|\hat{W} - \mathbf{W}\|$ on the right plot. For the three regularization schemes, the range of values with high probability of correct rank selection increases as n increases, and, most importantly achieves good mean squared error (right plot); in particular, for the non adaptive schemes (top plots), this corroborates the results from Proposition 9, which states that for $\lambda_n = \lambda_0 n^{-1/2}$ the probability tends to a limit in $(0, 1)$: indeed, when n increases, the value λ_n which achieves a particular limit grows as $n^{-1/2}$, and considering the log-scale for λ_n in Figure 3 and the uniform sampling for n in log-scale as well, the regular spacing between the decaying parts observed in Figure 3 is coherent with our results.

In Figure 4, we perform the same operations but with the inconsistent case of Figure 2. For the non adaptive case (top plot), the range of values of λ that achieve high probability of correct rank selection does not increase when n increases and stays bounded, in places where the estimation error is not tending to zero: in the inconsistent case, the trace norm regularization does not manage to solve the trade-off between rank consistency and regular consistency. However, for the adaptive versions, it does, still with a somewhat singular behavior of the limiting case $\gamma = 1$.

Finally, in Figure 5, we consider 400 different distributions with various values of $\|\Lambda\|_2$ smaller or greater than one, and computed the regularization paths with $n = 10^3$ samples. From the paths, we consider the estimate $\hat{W}$ with correct rank and best distance to $\mathbf{W}$ and plot the best error versus $\log_{10}(\|\Lambda\|_2)$. For positive values of $\log_{10}(\|\Lambda\|_2)$, the best error is far from zero, and the error grows with the distance to zero; while for negative values, we get low errors with lower errors for small $\log_{10}(\|\Lambda\|_2)$, corroborating the influence of $\|\Lambda\|_2$ described in Proposition 9.

7. Conclusion

We have presented an analysis of the rank consistency for the penalization by the trace norm, and derived general necessary and sufficient conditions. This work can be extended in several interesting ways: first, by going from the square loss to more general losses, in particular for other types of supervised learning problems such as classification; or by looking at the collaborative filtering setting where only some of the attributes are observed (Abernethy et al., 2006) and dimensions p and q are allowed to grow. Moreover, we are currently pursuing non asymptotic extensions of the current work, making links with the recent work of Recht et al. (2007) and of Meinshausen and Yu (2006).

Appendix A. Tools for Analysis of Singular Value Decomposition

In this appendix, we review and derive precise results regarding singular value decompositions. We consider $W \in \mathbb{R}^{p \times q}$ and we let denote $W = U \text{Diag}(s) V^\top$ its singular value decomposition with $U \in \mathbb{R}^{p \times r}$, $V \in \mathbb{R}^{q \times r}$ with orthonormal columns, and $s \in \mathbb{R}^r$ with strictly positive values (r is the rank
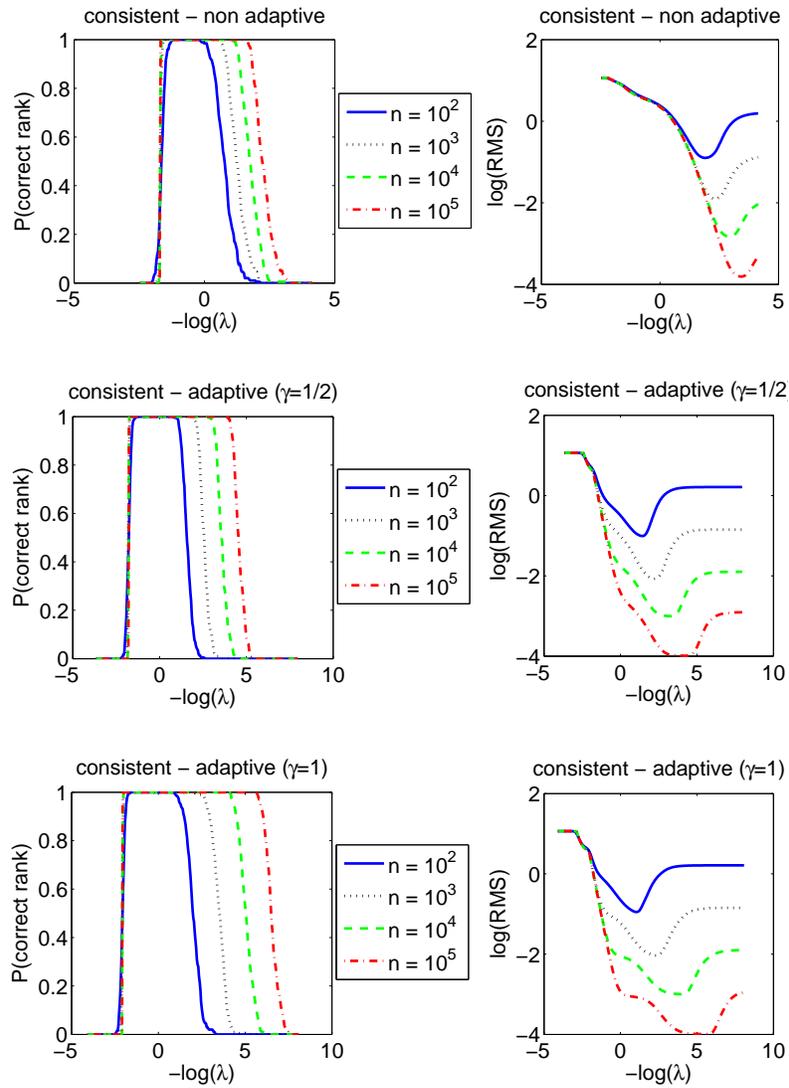
Figure 3: Synthetic example where consistency condition in Eq. (5) is satisfied: probability of correct rank selection (left) and logarithm of the expected mean squared estimation error (right), for several number of samples as a function of the regularization parameter, for regular regularization (top), adaptive regularization with $\gamma = 1/2$ (center) and $\gamma = 1$ (bottom).
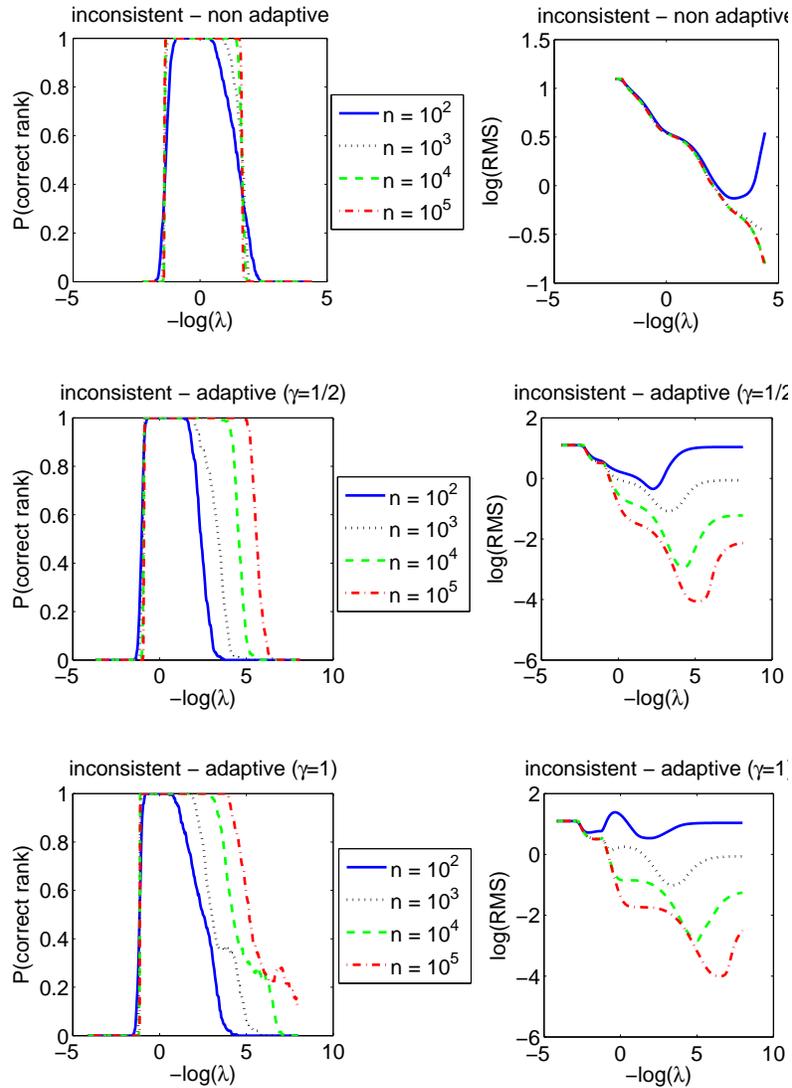
Figure 4: Synthetic example where consistency condition in Eq. (4) is not satisfied: probability of correct rank selection (left) and logarithm of the expected mean squared estimation error (right), for several number of samples as a function of the regularization parameter, for regular regularization (top), adaptive regularization with $\gamma = 1/2$ (center) and $\gamma = 1$ (bottom).
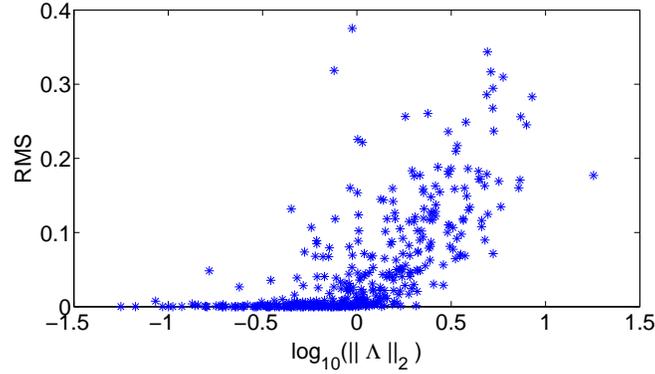
Figure 5: Scatter plots of $\log_{10}(\|\Lambda\|_2)$ versus the squared error of the best estimate with correct rank (i.e., such that $\text{rank}(\hat{W}) = \mathbf{r}$ and $\|\hat{W} - \mathbf{W}\|$ as small as possible). See text for details.

of W). Note that when a singular value s_i is simple, that is, does not coalesce with any other singular values, then the vectors u_i and v_i are uniquely defined up to simultaneous sign flips, that is, only the matrix $u_i v_i^\top$ is unique. However, when some singular values coalesce, then the corresponding singular vectors are defined up to a rotation, and thus in general care must be taken and considering isolated singular vectors should be avoided (Stewart and Sun, 1990). All tools presented in this appendix are robust to the particular choice of the singular vectors.

A.1 Jordan-Wielandt Matrix

We use the fact that singular values of W can be obtained from the eigenvalues of the Jordan-Wielandt matrix $\bar{W} = \begin{pmatrix} 0 & W \\ W^\top & 0 \end{pmatrix} \in \mathbb{R}^{(p+q)\times(p+q)}$ (Stewart and Sun, 1990). Indeed this matrix has eigenvalues s_i and $-s_i$, $i = 1, \ldots, r$, where s_i are the (strictly positive) singular values of W, with eigenvectors $\frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ v_i \end{pmatrix}$ and $\frac{1}{\sqrt{2}} \begin{pmatrix} u_i \\ -v_i \end{pmatrix}$ where u_i, v_i are the left and right associated singular vectors. Also, the remaining eigenvalues are all equal to zero, with eigensubspace (of dimension $p + q - 2r$) composed of all $\begin{pmatrix} u \\ v \end{pmatrix}$ such that for all $i \in \{1, \ldots, r\}$, $u^\top u_i = v^\top v_i = 0$. We let denote $\bar{U}$ the eigenvectors of $\bar{W}$ corresponding to non zero eigenvalues in $\bar{S}$. We have $\bar{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix}$ and $\bar{S} = \frac{1}{\sqrt{2}} \begin{pmatrix} \text{Diag}(s) & 0 \\ 0 & -\text{Diag}(s) \end{pmatrix}$ and $\bar{W} = \bar{U} \bar{S} \bar{U}^\top$, $\bar{U} \bar{U}^\top = \begin{pmatrix} UU^\top & 0 \\ 0 & VV^\top \end{pmatrix}$, and $\bar{U} \text{sign}(\bar{S}) \bar{U}^\top = \begin{pmatrix} 0 & UV^\top \\ VU^\top & 0 \end{pmatrix}$.

A.2 Cauchy Residue Formula and Eigenvalues

Given the matrix $\bar{W}$, and a simple closed curve $\mathcal{C}$ in the complex plane that does not go through any of the eigenvalues of $\bar{W}$, then

$$\Pi_C(\bar{W}) = \frac{1}{2i\pi} \oint_C \frac{d\lambda}{\lambda \mathbf{I} - \bar{W}}$$

is equal to the orthogonal projection onto the orthogonal sum of all eigensubspaces of $\bar{W}$ associated with eigenvalues in the interior of $\mathcal{C}$ (Kato, 1966). This is easily seen by writing down the eigenvalue decomposition and the Cauchy residue formula ($\frac{1}{2i\pi} \oint_{\mathcal{C}} \frac{d\lambda}{\lambda - \lambda_i} = 1$ if λ_i is in the interior $\text{int}(\mathcal{C})$ of $\mathcal{C}$ and 0 otherwise), and:

$$\frac{1}{2i\pi} \oint_{\mathcal{C}} \frac{d\lambda}{\lambda \mathbf{I} - \bar{W}} = \sum_{i=1}^{2r} \bar{u}_i \bar{u}_i^\top \times \frac{1}{2i\pi} \oint_{\mathcal{C}} \frac{d\lambda}{\lambda - \bar{s}_i} = \sum_{i,\ \bar{s}_i \in \text{int}(\mathcal{C})} u_i u_i^\top.$$

See Rudin (1987) for an introduction to complex analysis and Cauchy residue formula. Moreover, we can obtain the restriction of $\bar{W}$ onto a specific eigensubspace as:

$$\bar{W} \Pi_{\mathcal{C}}(\bar{W}) = \frac{1}{2i\pi} \oint_{\mathcal{C}} \frac{\bar{W} d\lambda}{\lambda \mathbf{I} - \bar{W}} = -\frac{1}{2i\pi} \oint_{\mathcal{C}} \frac{\lambda d\lambda}{\lambda \mathbf{I} - \bar{W}}.$$

We let denote s_1 and s_r the largest and smallest strictly positive singular values of W; if $\|\Delta\|_2 < s_r/2$, then $W + \Delta$ has r singular values strictly greater than $s_r/2$ and the remaining ones are strictly less than $s_r/2$ (Stewart and Sun, 1990). Thus, if we denote $\mathcal{C}$ the oriented circle of radius $s_r/2$, $\Pi_{\mathcal{C}}(\bar{W})$ is the projector on the $p + q - 2r$-dimensional null space of $\bar{W}$, and for any Δ such that $\|\Delta\|_2 < s_r/2$, $\Pi_{\mathcal{C}}(\bar{W} + \bar{\Delta})$ is also the projector on the $p + q - 2r$-dimensional invariant subspace of $\bar{W} + \bar{\Delta}$, which corresponds to the smallest eigenvalues. We let denote $\Pi_o(\bar{W} + \bar{\Delta})$ that projector and $\Pi_r(\bar{W} + \bar{\Delta}) = \mathbf{I} - \Pi_o(\bar{W} + \bar{\Delta})$ the orthogonal projector (which is the projection onto the $2r$-th principal subspace).

We can now find expansions around $\Delta = 0$ as follows:

$$
\begin{aligned}
\Pi_o(\bar{W} + \bar{\Delta}) - \Pi_o(\bar{W}) &= \frac{1}{2i\pi} \oint_{\mathcal{C}} (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W} - \bar{\Delta})^{-1} d\lambda \\
&= \frac{1}{2i\pi} \oint_{\mathcal{C}} (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W})^{-1} d\lambda \\
&\quad + \frac{1}{2i\pi} \oint_{\mathcal{C}} (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W} - \bar{\Delta})^{-1} d\lambda,
\end{aligned}
$$

and

$$
\begin{aligned}
(\bar{W} + \bar{\Delta}) \Pi_o(\bar{W} + \bar{\Delta}) - \bar{W} \Pi_o(\bar{W}) &= -\frac{1}{2i\pi} \oint_{\mathcal{C}} \lambda (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W} - \bar{\Delta})^{-1} d\lambda \\
&= -\frac{1}{2i\pi} \oint_{\mathcal{C}} \lambda (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W})^{-1} d\lambda \\
&\quad - \frac{1}{2i\pi} \oint_{\mathcal{C}} \lambda (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda \mathbf{I} - \bar{W} - \bar{\Delta})^{-1} d\lambda,
\end{aligned}
$$

which lead to the following two propositions:

Proposition 16 *Assume W has rank r and $\|\Delta\|_2 < s_r/4$ where s_r is the smallest positive singular value of W. Then the projection $\Pi_r(\bar{W})$ on the first r eigenvectors of $\bar{W}$ is such that*

$$\|\Pi_o(\bar{W} + \bar{\Delta}) - \Pi_o(\bar{W})\|_2 \leqslant \frac{4}{s_r} \|\Delta\|_2$$

and

$$\|\Pi_o(\bar{W} + \bar{\Delta}) - \Pi_o(\bar{W}) - (\mathbf{I} - \bar{U}\bar{U}^\top)\bar{\Delta}\bar{U}\bar{S}^{-1}\bar{U}^\top - \bar{U}\bar{S}^{-1}\bar{U}^\top\bar{\Delta}(\mathbf{I} - \bar{U}\bar{U}^\top)\|_2 \leqslant \frac{8}{s_r^2} \|\bar{\Delta}\|_2^2.$$

Proof For $\lambda \in C$ we have: $\|(\lambda\mathbf{I} - \bar{W})^{-1}\|_2 \geqslant 2/s_r$ and $\|(\lambda\mathbf{I} - \bar{W} - \bar{\Delta})^{-1}\|_2 \geqslant 4/s_r$, which implies

$$
\begin{aligned}
\|\Pi_r(\bar{W} + \bar{\Delta}) - \Pi_r(\bar{W})\|_2 &\leqslant \frac{1}{2\pi} \oint_C \|(\lambda\mathbf{I} - \bar{W})^{-1}\|_2 \|\Delta\|_2 \|(\lambda\mathbf{I} - \bar{W} - \bar{\Delta})^{-1}\|_2 \\
&\leqslant \left(\frac{1}{2\pi} 2\pi \frac{s_r}{2} \right) \|\Delta\|_2 \frac{2}{s_r} \frac{4}{s_r}.
\end{aligned}
$$

In order to prove the other result, we simply need to compute:

$$
\begin{aligned}
\frac{1}{2i\pi} \oint_C (\lambda\mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda\mathbf{I} - \bar{W})^{-1} d\lambda &= \sum_{i,j} \bar{u}_i \bar{u}_i^\top \Delta \bar{v}_j \bar{v}_j^\top \frac{1}{2i\pi} \oint_C \frac{1}{(\lambda - \bar{s}_i)(\lambda - \bar{s}_j)} d\lambda \\
&= \sum_{i,j} \bar{u}_i \bar{u}_i^\top \Delta \bar{v}_j \bar{v}_j^\top \left(\frac{1_{i\notin\mathrm{int}(C)} 1_{j\in\mathrm{int}(C)}}{\bar{s}_j} + \frac{1_{j\notin\mathrm{int}(C)} 1_{i\in\mathrm{int}(C)}}{\bar{s}_i} \right) \\
&= (\mathbf{I} - \bar{U}\bar{U}^\top) \bar{\Delta} \bar{U} \bar{S}^{-1} \bar{U}^\top + \bar{U} \bar{S}^{-1} \bar{U}^\top \bar{\Delta} (\mathbf{I} - \bar{U}\bar{U}^\top).
\end{aligned}
$$

∎

Proposition 17 *Assume W has rank r and $\|\Delta\|_2 < s_r/4$ where s_r is the smallest positive singular value of W. Then the projection $\Pi_r(\bar{W})$ on the first r eigenvectors of $\bar{W}$ is such that*

$$
\|\Pi_o(\bar{W} + \bar{\Delta})(\bar{W} + \bar{\Delta}) - \Pi_o(\bar{W})\bar{W}\|_2 \leqslant 2\|\Delta\|_2
$$

and

$$
\|\Pi_o(\bar{W} + \bar{\Delta})(\bar{W} + \bar{\Delta}) - \Pi_o(\bar{W})\bar{W} + (\mathbf{I} - \bar{U}\bar{U}^\top) \bar{\Delta} (\mathbf{I} - \bar{U}\bar{U}^\top)\|_2 \leqslant \frac{4}{s_r} \|\bar{\Delta}\|_2^2.
$$

Proof For $\lambda \in C$ we have: $\|(\lambda\mathbf{I} - \bar{W})^{-1}\|_2 \geqslant 2/s_r$ and $\|(\lambda\mathbf{I} - \bar{W} - \bar{\Delta})^{-1}\|_2 \geqslant 4/s_r$, which implies

$$
\begin{aligned}
\|\Pi_r(\bar{W} + \bar{\Delta}) - \Pi_r(\bar{W})\|_2 &\leqslant \frac{1}{2\pi} \oint_C |\lambda| \|(\lambda\mathbf{I} - \bar{W})^{-1}\|_2 \|\Delta\|_2 \|(\lambda\mathbf{I} - \bar{W} - \bar{\Delta})^{-1}\|_2 \\
&\leqslant \left(\frac{1}{2\pi} 2\pi \frac{s_r}{2} \right) \frac{s_r}{2} \|\Delta\|_2 \frac{2}{s_r} \frac{4}{s_r}.
\end{aligned}
$$

In order to prove the other result, we simply need to compute:

$$
\begin{aligned}
-\frac{1}{2i\pi} \oint_C \lambda (\lambda\mathbf{I} - \bar{W})^{-1} \bar{\Delta} (\lambda\mathbf{I} - \bar{W})^{-1} d\lambda &= -\sum_{i,j} \bar{u}_i \bar{u}_i^\top \Delta \bar{v}_j \bar{v}_j^\top \frac{1}{2i\pi} \oint_C \frac{\lambda}{(\lambda - \bar{s}_i)(\lambda - \bar{s}_j)} d\lambda \\
&= -\sum_{i,j} \bar{u}_i \bar{u}_i^\top \Delta \bar{v}_j \bar{v}_j^\top \left(1_{i\in\mathrm{int}(C)} 1_{j\in\mathrm{int}(C)} \right) \\
&= -(\mathbf{I} - \bar{U}\bar{U}^\top) \bar{\Delta} (\mathbf{I} - \bar{U}\bar{U}^\top).
\end{aligned}
$$

∎

The variations of $\Pi(\bar{W})$ translates immediately into variations of the singular projections $UU^\top$ and $VV^\top$. Indeed we get that the first order variation of $UU^\top$ is $-(\mathbf{I} - UU^\top)\Delta V S^{-1} U^\top$ and the variation of V is equal to $-(\mathbf{I} - VV^\top)\Delta^\top U S^{-1} V^\top$, with errors bounded in spectral norm by $\frac{8}{s_r^2}\|\Delta\|_2^2$. Similarly, when restricted to the small singular values, the first order expansion is $(\mathbf{I} - UU^\top)\Delta(\mathbf{I} - VV^\top)$, with error term bounded in spectral norm by $\frac{4}{s_r}\|\Delta\|_2^2$. Those results lead to the following proposition that gives a local sufficient condition for $\mathrm{rank}(W + \Delta) > \mathrm{rank}(W)$:

Proposition 18 *Assume W has rank $r < \min\{p,q\}$ with ordered singular value decomposition $W = U\operatorname{Diag}(s)V^\top$. If $\frac{4}{s_r}\|\Delta\|_2^2 < \|(\mathbf{I}-UU^\top)\Delta(\mathbf{I}-VV^\top)\|_2$, then $\operatorname{rank}(W+\Delta) > r$.*

Appendix B. Some Facts about the Trace Norm

In this appendix, we review known properties of the trace norm that we use in this paper. Most of the results are extensions of similar results for the ℓ_1-norm on vectors. First, we have the following result:

Lemma 19 (Dual norm, Fazel et al., 2001) *The trace norm $\|\cdot\|_*$ is a norm and its dual norm is the operator norm $\|\cdot\|$.*

Note that the dual norm $N(W)$ is defined as Boyd and Vandenberghe (2003):

$$N(W) = \sup_{\|V\|_* \leqslant 1} \operatorname{tr} W^\top V.$$

This immediately implies the following result:

Lemma 20 (Fenchel conjugate) *We have:* $\max_{W \in \mathbb{R}^{p\times q}} \operatorname{tr} W^\top V - \|W\|_* = 0$ *if $\|V\| \leqslant 1$ and $+\infty$ otherwise.*

In this paper, we need to compute the subdifferential and directional derivatives of the trace norm. We have from Recht et al. (2007) or Borwein and Lewis (2000):

Proposition 21 (Subdifferential) *If $W = U\operatorname{Diag}(s)V^\top$ with $U \in \mathbb{R}^{p\times m}$ and $V \in \mathbb{R}^{q\times m}$ having orthonormal columns, and $s \in \mathbb{R}^m$ is strictly positive, is the singular value decomposition of W, then $\|W\|_* = \sum_{i=1}^{m} s_i$ and the subdifferential of $\|\cdot\|_*$ is equal to*

$$\partial\|\cdot\|_*(W) = \left\{ UV^\top + M, \text{ such that } \|M\|_2 \leqslant 1, \ U^\top M = 0 \text{ and } MV = 0 \right\}.$$

This result can be extended to compute directional derivatives:

Proposition 22 (Directional derivative) *The directional derivative at $W = USV^\top$ is equal to:*

$$\lim_{\varepsilon \to 0^+} \frac{\|W+\varepsilon\Delta\|_* - \|W\|_*}{\varepsilon} = \operatorname{tr} U^\top \Delta V + \|U_\perp^\top \Delta V_\perp\|_*,$$

where $U_\perp \in \mathbb{R}^{p\times(p-m)}$ and $V_\perp \in \mathbb{R}^{q\times(q-m)}$ are any orthonormal complements of U and V.

Proof From the subdifferential, we get the directional derivative (Borwein and Lewis, 2000) as

$$\lim_{\varepsilon \to 0^+} \frac{\|W+\varepsilon\Delta\|_* - \|W\|_*}{\varepsilon} = \max_{V \in \partial\|\cdot\|_*(W)} \operatorname{tr} \Delta^\top V$$

which exactly leads to the desired result. ∎

The final result that we use is a bit finer as it gives an upper bound on the error in the previous limit:

Proposition 23 *Let $W = U \operatorname{Diag}(s) V^\top$ the ordered singular value decomposition, where $\operatorname{rank}(W) = r$, $s > 0$ and $U_\perp$ and $V_\perp$ be orthogonal complement of U and V; then, if $\|\Delta\|_2 \leqslant s_r/4$:*

$$\left| \|W + \Delta\|_* - \|W\|_* - \operatorname{tr} U^\top \Delta V - \|U_\perp^\top \Delta V_\perp\|_* \right| \leqslant 16 \min\{p, q\} \frac{s_1^2}{s_r^3} \|\Delta\|_2^2.$$

Proof The trace norm of $\|W + \Delta\|_*$ may be divided into the sum of the r largest and the sum of the remaining singular values. The sums of the remaining ones are given through Proposition 17 by $\|U_\perp^\top \Delta V_\perp\|_*$ with an error bounded by $\min\{p, q\} \frac{4}{s_r} \|\Delta\|_2^2$. For the first r singular values, we need to upperbound the second derivative of the sum of the r largest eigenvalues of $\bar{W} + \bar{\Delta}$ with strictly positive eigengap, which leads to the given bound by using the same Cauchy residue technique described in Appendix A. ∎

Appendix C. Proofs

In this appendix, we give the proofs of the results presented in the paper.

C.1 Proof of Lemma 2

We let denote $S \in \{0, 1\}^{n_x \times n_y}$ the sampling matrix; that is, $S_{ij} = 1$ if the pair (i, j) is observed and zero otherwise. We let denote $\tilde{X}$ and $\tilde{Y}$ the data matrices. We can write $M_k = \tilde{X}^\top \delta_{i_k} \delta_{j_k}^\top \tilde{Y}$ and:

$$\frac{1}{n} \sum_{k=1}^n \operatorname{vec}(M_k) \operatorname{vec}(M_k)^\top = \frac{1}{n} \sum_{k=1}^n (\tilde{Y} \otimes \tilde{X})^\top \operatorname{vec}(\delta_{i_k} \delta_{j_k}^\top) \operatorname{vec}(\delta_{i_k} \delta_{j_k}^\top)^\top (\tilde{Y} \otimes \tilde{X})$$

$$= \frac{1}{n} (\tilde{Y} \otimes \tilde{X})^\top \operatorname{Diag}(\operatorname{vec}(S))(\tilde{Y} \otimes \tilde{X}),$$

which leads to (denoting $\hat{\Sigma}_{xx} = n_x^{-1} \tilde{X}^\top \tilde{X}$ and $\hat{\Sigma}_{yy} = n_x^{-1} \tilde{Y}^\top \tilde{Y}$):

$$\left(\frac{1}{n} \sum_{k=1}^n \operatorname{vec}(M_k) \operatorname{vec}(M_k)^\top - \hat{\Sigma}_{yy} \otimes \hat{\Sigma}_{xx} \right) = \frac{1}{n} (\tilde{Y} \otimes \tilde{X})^\top \operatorname{Diag}(\operatorname{vec}(S - n/n_x n_y))(\tilde{Y} \otimes \tilde{X}).$$

We can thus compute the squared Frobenius norm:

$$\left\| \frac{1}{n} \sum_{k=1}^n \operatorname{vec}(M_k) \operatorname{vec}(M_k)^\top - \hat{\Sigma}_{yy} \otimes \hat{\Sigma}_{xx} \right\|_F^2$$

$$= \frac{1}{n^2} \operatorname{tr} \operatorname{Diag}(\operatorname{vec}(S - n/n_x n_y))(\tilde{Y}\tilde{Y}^\top \otimes \tilde{X}\tilde{X}^\top) \operatorname{Diag}(\operatorname{vec}(S - n/n_x n_y))(\tilde{Y}\tilde{Y}^\top \otimes \tilde{X}\tilde{X}^\top)$$

$$= \frac{1}{n^2} \sum_{i,j,i',j'} (S_{ij} - n/n_x n_y)(\tilde{Y}\tilde{Y}^\top \otimes \tilde{X}\tilde{X}^\top)_{ij,i'j'} (S_{i'j'} - n/n_x n_y)(\tilde{Y}\tilde{Y}^\top \otimes \tilde{X}\tilde{X}^\top)_{ij,i'j'}.$$

We have, by properties of sampling without replacement (Hoeffding, 1963):

$$\mathbb{E}(S_{ij} - n/n_x n_y)(S_{i'j'} - n/n_x n_y) = n/n_x n_y (1 - n/n_x n_y) \text{ if } (i, j) = (i', j'),$$

$$\mathbb{E}(S_{ij} - n/n_x n_y)(S_{i'j'} - n/n_x n_y) = -n/n_x n_y (1 - n/n_x n_y) \frac{1}{n_x n_y - 1} \text{ if } (i, j) \neq (i', j').$$

This implies

$$\mathbb{E}(\|\frac{1}{n}\sum_{k=1}^{n}\mathrm{vec}(M_k)\,\mathrm{vec}(M_k)^{\top} - \hat{\Sigma}_{yy}\otimes\hat{\Sigma}_{xx}\|_F^2|\tilde{X},\tilde{Y})$$

$$= \frac{1}{n_x n_y n}\sum_{i,j}(\tilde{Y}\tilde{Y}^{\top}\otimes\tilde{X}\tilde{X}^{\top})_{ij,ij}^2 - \frac{1}{(n_x n_y - 1)n_x n_y n}\sum_{(i,j)\neq(i',j')}(\tilde{Y}\tilde{Y}^{\top}\otimes\tilde{X}\tilde{X}^{\top})_{ij,i'j'}^2$$

$$\leqslant \frac{2}{n_x n_y n}\sum_{i,j}\|\tilde{y}_j\|^4\|\tilde{x}_i\|^4.$$

This finally implies that

$$\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^{n}\mathrm{vec}(M_k)\,\mathrm{vec}(M_k)^{\top} - \Sigma_{yy}\otimes\Sigma_{xx}\right\|_F^2$$

$$\leqslant \frac{4}{n}\sum_{i,j}\mathbb{E}\|x\|^4\mathbb{E}\|y\|^4 + 2\mathbb{E}\|\hat{\Sigma}_{xx}-\Sigma_{xx}\|_F^2\mathbb{E}\|\hat{\Sigma}_{yy}\|_F^2 + 2\mathbb{E}\|\hat{\Sigma}_{yy}-\Sigma_{yy}\|_F^2\|\Sigma_{xx}\|_F^2$$

$$\leqslant C\mathbb{E}\|x\|^4\mathbb{E}\|y\|^4 \times (\frac{1}{n}+\frac{1}{n_y}+\frac{1}{n_x}),$$

for some constant $C > 0$. This implies (**A2**). To prove the asymptotic normality in (**A3**), we use the martingale central limit theorem (Hall and Heyde, 1980) with sequence of σ-fields $\mathcal{F}_{n,k} = \sigma(\tilde{X},\tilde{Y},\varepsilon_1,\dots,\varepsilon_k,(i_1,j_1),\dots,(i_k,j_k))$ for $k \leqslant n$. We consider $\Delta_{n,k} = n^{-1/2}\varepsilon_{i_k j_k}y_{j_k}\otimes x_{i_k} \in \mathbb{R}^{pq}$ as the martingale difference. We have $\mathbb{E}(\Delta_{n,k}|\mathcal{F}_{n,k-1}) = 0$ and

$$\mathbb{E}(\Delta_{n,k}\Delta_{n,k}^{\top}|\mathcal{F}_{n,k-1}) = n^{-1}\sigma^2 y_{j_k}y_{j_k}^{\top}\otimes x_{i_k}x_{i_k}^{\top},$$

with $\mathbb{E}(\|\Delta_{n,k}\|^4) = O(n^{-2})$ because of the finite fourth order moments. Moreover,

$$\sum_{k=1}^{n}\mathbb{E}(\Delta_{n,k}\Delta_{n,k}^{\top}|\mathcal{F}_{n,k-1}) = \sigma^2\hat{\Sigma}_{mm},$$

and thus tends in probability to $\sigma^2\Sigma_{yy}\otimes\Sigma_{xx}$ because of (**A2**). The assumptions of the martingale central limit theorem are met, we have that $\sum_{k=1}^{n}\mathrm{vec}(\Delta_{n,k})$ is asymptotically normal with mean zero and covariance matrix $\sigma^2\Sigma_{yy}\otimes\Sigma_{xx}$, which concludes the proof.

C.2 Proof of Proposition 4

We may first restrict minimization over the ball $\{W, \|W\|_* \leqslant \|\hat{\Sigma}_{mm}^{-1}\hat{\Sigma}_{Mz}\|_*\}$ because the optimum value is less than the value for $W = \hat{\Sigma}_{mm}^{-1}\hat{\Sigma}_{Mz}$. Since this random variable is bounded in probability, we can reduce the problem to a compact set. The sequence of continuous random functions $W \mapsto \frac{1}{2}\mathrm{vec}(W-\mathbf{W})^{\top}\hat{\Sigma}_{mm}\mathrm{vec}(W-\mathbf{W}) - \mathrm{tr}W^{\top}\hat{\Sigma}_{M\varepsilon} + \lambda_n\|W\|_*$ converges pointwise in probability to $W \mapsto \frac{1}{2}\mathrm{vec}(W-\mathbf{W})^{\top}\Sigma_{mm}\mathrm{vec}(W-\mathbf{W}) + \lambda_0\|W\|_*$ with a unique global minimum (because Σ_{mm} is assumed invertible). We can thus apply standard result of consistency in M-estimation (Van der Vaart, 1998; Shao, 2003).

C.3 Proof of Proposition 7

We consider the result of Proposition 6: $\hat{\Delta} = n^{1/2}(\hat{W} - W)$ is asymptotically normal with mean zero and covariance $\sigma^2 \Sigma_{mm}^{-1}$. By Proposition 18 in Appendix B, if $\frac{4n^{-1/2}}{s_r}\|\hat{\Delta}\|_2^2 < \|U_\perp^\top \hat{\Delta} V_\perp\|_2$, then rank$(\hat{W}) > \mathbf{r}$. For a random variable Θ with normal distribution with mean zero and covariance matrix $\sigma^2 \Sigma_{mm}^{-1}$, we let denote $f(C) = \mathbb{P}(\frac{4C^{-1/2}}{s_r}\|\Theta\|_2^2 < \|U_\perp^\top \Theta V_\perp\|_2)$. By the dominated convergence theorem, $f(C)$ converges to one when $C \to \infty$. Let $\varepsilon > 0$, thus there exists $C_0 > 0$ such that $f(C_0) > 1 - \varepsilon/2$. By the asymptotic normality result, $\mathbb{P}(\frac{4C_0^{-1/2}}{s_r}\|\hat{\Delta}\|_2^2 < \|U_\perp^\top \hat{\Delta} V_\perp\|_2)$ converges to $f(C_0)$ thus $\exists n_0 > 0$ such that $\forall n > n_0$, $\mathbb{P}(\frac{4C_0^{-1/2}}{s_r}\|\hat{\Delta}\|_2^2 < \|U_\perp^\top \hat{\Delta} V_\perp\|_2) > f(C_0) - \varepsilon/2 > 1 - \varepsilon$, which concludes the proof, because $\mathbb{P}(\frac{4n^{-1/2}}{s_r}\|\hat{\Delta}\|_2^2 < \|U_\perp^\top \hat{\Delta} V_\perp\|_2) \geqslant \mathbb{P}(\frac{4C_0^{-1/2}}{s_r}\|\hat{\Delta}\|_2^2 < \|U_\perp^\top \hat{\Delta} V_\perp\|_2)$ as soon as $n > C_0$.

C.4 Proof of Proposition 8

This is the same result as Fu and Knight (2000), but extended to the trace norm minimization, simply using the directional derivative result of Proposition 22 and the epiconvergence theorem from Geyer (1994, 1996). Indeed, if we denote $V_n(\Delta) = \text{vec}(\Delta)^\top \hat{\Sigma}_{mm} \text{vec}(\Delta) - \text{tr} \Delta^\top n^{1/2} \hat{\Sigma}_{M\varepsilon} + \lambda_0 n^{1/2}(\|W + n^{-1/2}\Delta\|_* - \|W\|_*)$ and $V(\Delta) = \text{vec}(\Delta)^\top \Sigma_{mm} \text{vec}(\Delta) - \text{tr} \Delta^\top A + \lambda_0 [\text{tr} U^\top \Delta V + \|U_\perp^\top \Delta V_\perp\|_*]$, then for each Δ, $V_n(\Delta)$ converges in probability to $V(\Delta)$, and V is strictly convex, which implies that it has an unique global minimum; thus the epi-convergence theorem can be applied, which concludes the proof.

Note that a simpler analysis using regular tools in M-estimation leads to $\hat{W} = W + n^{-1/2}\hat{\Delta} + o_p(n^{-1/2})$, where $\hat{\Delta}$ is the unique global minimizer of

$$\min_{\Delta \in \mathbb{R}^{p \times q}} \frac{1}{2} \text{vec}(\Delta)^\top \Sigma_{mm} \text{vec}(\Delta) - \text{tr} \Delta^\top (n^{1/2}\hat{\Sigma}_{M\varepsilon}) + \lambda_0 \left[\text{tr} U^\top \Delta V + \|U_\perp^\top \Delta V_\perp\|_* \right],$$

that is, we can actually take $A = n^{1/2}\hat{\Sigma}_{M\varepsilon}$ (which is asymptotically normal with correct moments).

C.5 Proof of Proposition 9

We let denote $\hat{\Delta} = n^{1/2}(\hat{W} - W)$. We first show that $\limsup_{n \to \infty} \mathbb{P}(\text{rank}(\hat{W}) = \mathbf{r})$ is smaller than the proposed limit a. We consider the following events:

$$
\begin{aligned}
E_0 &= \{\text{rank}(\hat{W}) = \mathbf{r}\} \\
E_1 &= \{\|n^{-1/2}\hat{\Delta}\|_2 < \mathbf{s}_r/2\} \\
E_2 &= \left\{ \frac{4n^{-1/2}}{s_r}\|\hat{\Delta}\|_2^2 < \|U_\perp^\top \hat{\Delta} V_\perp\|_2 \right\}.
\end{aligned}
$$

By Proposition 18 in Appendix B, we have $E_1 \cap E_2 \subset E_0^c$, and thus it suffices to show that $\mathbb{P}(E_1)$ tends to one, while $\limsup_{n \to \infty} \mathbb{P}(E_2^c) \leqslant a$. The first assertion is a simple consequence of Proposition 8.

Moreover, by Proposition 8, $\hat{\Delta}$ converges in distribution to the unique global optimum $\Delta(A)$ of an optimization problem parameterized by a vector A with normal distribution. For a given $\eta > 0$, we consider the probability $\mathbb{P}(\|U_\perp^\top \Delta(A) V_\perp\|_2 \leqslant \eta)$. For any A, when η tends to zero, the indicator function $1_{\|U_\perp^\top \Delta(A) V_\perp\|_2 \leqslant \eta}$ converges to $1_{\|U_\perp^\top \Delta(A) V_\perp\|_2 = 0}$, which is equal to $1_{\|\Lambda(A)\|_2 \leqslant \lambda_0}$, where

$$\text{vec}(\Lambda(A)) = \left((V_\perp \otimes U_\perp)^\top \Sigma_{mm}^{-1} (V_\perp \otimes U_\perp) \right)^{-1} \left((V_\perp \otimes U_\perp)^\top \Sigma_{mm}^{-1} ((V \otimes U)\text{vec}(I) - \text{vec}(A)) \right).$$

By the dominated convergence theorem, $\mathbb{P}(\|\mathbf{U}_\perp^\top \Delta(A)\mathbf{V}_\perp\|_2 \leqslant \eta)$ converges to

$$a = \mathbb{P}(\|\Lambda(A)\|_2 \leqslant \lambda_0),$$

which is the proposed limit. This limit is in $(0,1)$ because of the normal distribution has an invertible covariance matrix and the set $\{\|\Lambda\|_2 \leqslant 1\}$ and its complement have non empty interiors.

Since $\hat{\Delta} = O_p(1)$, we can instead consider $E_3 = \{\frac{4n^{-1/2}}{s_r}M^2 < \|\mathbf{U}_o^\top \hat{\Delta}\mathbf{V}_\perp\|_2\}$ for a particular M, instead of E_2. Then following the same line or arguments than in Appendix C.3, we conclude that $\limsup_{n\to\infty} \mathbb{P}(E_3^c) \leqslant a$, which concludes the first part of the proof.

We now show that $\liminf_{n\to\infty} \mathbb{P}(\text{rank}(\hat{W}) = \mathbf{r}) \geqslant a$. A sufficient condition for rank consistency is the following: we let denote $\hat{W} = USV^\top$ the singular value decomposition of $\hat{W}$ and we let denote U_o and V_o the singular vectors corresponding to all but the $\mathbf{r}$ largest singular values. Since we have simultaneous singular value decompositions, a sufficient condition is that $\text{rank}(\hat{W}) \geqslant \mathbf{r}$ and $\left\|U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon}\right)V_o\right\|_2 < \lambda_n(1-\eta)$. If $\|\Lambda(n^{1/2}\hat{\Sigma}_{M\varepsilon})\| \leqslant \lambda_0(1-\eta)$, then, by Lemma 11, $\mathbf{U}_\perp^\top \Delta(n^{1/2}\hat{\Sigma}_{M\varepsilon})\mathbf{V}_\perp = 0$, and we get, using the proof of Proposition 8 and the notation $\hat{A} = n^{1/2}\hat{\Sigma}_{M\varepsilon}$:

$$U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon}\right)V_o = U_o^\top n^{-1/2}\left(\hat{\Sigma}_{mm}\Delta(\hat{A}) - \hat{A}\right)V_o + o_p(n^{-1/2}).$$

Moreover, because of regular consistency and a positive eigengap for $\mathbf{W}$, the projection onto the first $\mathbf{r}$ singular vectors of $\hat{W}$ converges to the projection onto the first $\mathbf{r}$ singular vectors of $\mathbf{W}$ (see Appendix A), which implies that the projection onto the orthogonal is also consistent, that is, $U_o U_o^\top$ converges in probability to $\mathbf{U}_\perp \mathbf{U}_\perp^\top$ and $V_o V_o^\top$ converges in probability to $\mathbf{V}_\perp \mathbf{V}_\perp^\top$. Thus:

$$\left\|U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon}\right)V_o\right\|_2 = \left\|U_o U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon}\right)V_o V_o^\top\right\|_2$$
$$= n^{-1/2}\|\mathbf{U}_\perp \mathbf{U}_\perp^\top(\hat{\Sigma}_{mm}\Delta(\hat{A}) - \hat{A})\mathbf{V}_\perp \mathbf{V}_\perp^\top\|_2 + o_p(n^{-1/2})$$
$$= n^{-1/2}\|\Lambda(A)\|_2 + o_p(n^{-1/2}).$$

This implies that

$$\lim_{n\to\infty}\inf \left\|U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon}\right)V_o\right\|_2 < \lambda_n(1-\eta) \geqslant \lim_{n\to\infty}\inf \mathbb{P}(\|\Lambda(\hat{A})\|_2 \leqslant \lambda_0(1-\eta))$$

which converges to a when η tends to zero, which concludes the proof.

C.6 Proof of Proposition 10

This is the same result as Fu and Knight (2000), but extended to the trace norm minimization, simply using the directional derivative result of Proposition 22. If we write $\hat{W} = \mathbf{W} + \lambda_n \hat{\Delta}$, then $\hat{\Delta}$ is defined as the global minimum of

$$V_n(\Delta) = \frac{1}{2}\text{vec}(\Delta)^\top \hat{\Sigma}_{mm}\text{vec}(\Delta) - \lambda_n^{-1}\text{tr}\Delta^\top \hat{\Sigma}_{M\varepsilon} + \lambda_n^{-1}(\|\mathbf{W} + \lambda_n\Delta\|_* - \|\mathbf{W}\|_*)$$
$$= \frac{1}{2}\text{vec}(\Delta)^\top \Sigma_{mm}\text{vec}(\Delta) + O_p(\zeta_n\|\Delta\|_2^2) + O_p(\lambda_n^{-1}n^{-1/2}) + \text{tr}\Delta^\top \hat{\Sigma}_{M\varepsilon}$$
$$+ \text{tr}\mathbf{U}^\top \Delta\mathbf{V} + \|\mathbf{U}_\perp^\top \Delta\mathbf{V}_\perp\|_* + O_p(\lambda_n\|\Delta\|_2^2)$$
$$= V(\Delta) + O_p(\zeta_n\|\Delta\|_2^2) + O_p(\lambda_n^{-1}n^{-1/2}) + O_p(\lambda_n\|\Delta\|_2^2).$$

More precisely, if $M\lambda_n < \mathbf{s}_r/2$,

$$
\begin{aligned}
\mathbb{E} \sup_{\|\Delta\|_2 \leqslant M} |V_n(\Delta) - V(\Delta)| &= \text{cst} \times \left(M^2 \mathbb{E}\|\hat{\Sigma}_{mm} - \Sigma_{mm}\|_F + M\lambda_n^{-1}\mathbb{E}(\|\hat{\Sigma}_{M\varepsilon}\|^2)^{1/2} + \lambda_n M^2 \right) \\
&= O(M^2\zeta_n + M\lambda_n^{-1}n^{-1/2} + \lambda_n M^2).
\end{aligned}
$$

Moreover, $V(\Delta)$ achieves its minimum at a bounded point $\Delta_0 \neq 0$. Thus, by Markov inequality the minimum of $V_n(\Delta)$ over the ball $\|\Delta\|_2 < 2\|\Delta_0\|_2$ is with probability tending to one strictly inside and is thus also the unconstrained minimum, which leads to the proposition.

C.7 Proof of Proposition 11

The optimal $\Delta \in \mathbb{R}^{p\times q}$ should be such that $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp$ has low rank, where $\mathbf{U}_\perp \in \mathbb{R}^{p\times(p-\mathbf{r})}$ and $\mathbf{V}_\perp \in \mathbb{R}^{q\times(q-\mathbf{r})}$ are orthogonal complements of the singular vectors $\mathbf{U}$ and $\mathbf{V}$. We now derive the condition under which the optimal Δ is such that $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp$ is actually equal to zero: we consider the minimum of $\frac{1}{2}\text{vec}(\Delta)^\top \Sigma_{mm}\text{vec}(\Delta) + \text{vec}(\Delta)^\top \text{vec}(\mathbf{U}\mathbf{V}^\top)$ with respect to Δ such that $\text{vec}(\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp) = (\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \text{vec}(\Delta) = 0$. The solution of that constrained optimization problem is obtained through the following linear system (Boyd and Vandenberghe, 2003):

$$
\begin{pmatrix} \Sigma_{mm} & (\mathbf{V}_\perp \otimes \mathbf{U}_\perp) \\ (\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top & 0 \end{pmatrix} \begin{pmatrix} \text{vec}(\Delta) \\ \text{vec}(\Lambda) \end{pmatrix} = \begin{pmatrix} -\text{vec}(\mathbf{U}\mathbf{V}^\top) \\ 0 \end{pmatrix},
$$

where $\Lambda \in \mathbb{R}^{(p-\mathbf{r})\times(q-\mathbf{r})}$ is the Lagrange multiplier for the equality constraint. We can solve explicitly for Δ and Λ which leads to

$$
\text{vec}(\Lambda) = \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1}(\mathbf{V}_\perp \otimes \mathbf{U}_\perp) \right)^{-1} \left((\mathbf{V}_\perp \otimes \mathbf{U}_\perp)^\top \Sigma_{mm}^{-1}(\mathbf{V} \otimes \mathbf{U})\text{vec}(\mathbf{I}) \right),
$$

and

$$
\text{vec}(\Delta) = -\Sigma_{mm}^{-1}\text{vec}(\mathbf{U}\mathbf{V}^\top - \mathbf{U}_\perp \Lambda \mathbf{V}_\perp^\top).
$$

Then the minimum of the function $F(\Delta)$ in Eq. (2) is such that $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp = 0$ (and thus equal to Δ defined above) if and only if for all $\Theta \in \mathbb{R}^{p\times q}$, the directional derivative of F at Δ in the direction Θ is nonnegative, that is:

$$
\lim_{\varepsilon \to 0^+} \frac{F(\Delta + \varepsilon\Theta) - F(\Delta)}{\varepsilon} \geqslant 0.
$$

By Proposition 22, this directional derivative is equal to

$$
\begin{aligned}
\text{tr}\Theta^\top(\Sigma_{mm}\Delta + \mathbf{U}\mathbf{V}^\top) + \|\mathbf{U}_\perp^\top\Theta\mathbf{V}_\perp\|_* &= \text{tr}\Theta^\top\mathbf{U}_\perp\Lambda\mathbf{V}_\perp + \|\mathbf{U}_\perp^\top\Theta\mathbf{V}_\perp\|_* \\
&= \text{tr}\Lambda^\top\mathbf{U}_\perp^\top\Theta\mathbf{V}_\perp + \|\mathbf{U}_\perp^\top\Theta\mathbf{V}_\perp\|_*.
\end{aligned}
$$

Thus the directional derivative is always non negative if for all $\Theta' \in \mathbb{R}^{(p-\mathbf{r})\times(q-\mathbf{r})}$, $\text{tr}\Lambda^\top\Theta' + \|\Theta'\|_* \geqslant 0$, that is, if and only if $\|\Lambda\|_2 \leqslant 1$, which concludes the proof.

C.8 Proof of Theorem 12

Regular consistency is obtained by Corollary 5. We consider the problem in Eq. (2) of Proposition 10, where $\lambda_n n^{1/2} \to \infty$ and $\lambda_n \to 0$. Since Eq. (5) is satisfied, the solution Δ indeed satisfies $\mathbf{U}_\perp^\top \Delta \mathbf{V}_\perp = 0$ by Lemma 11.

We have $\hat{W} = \mathbf{W} + \lambda_n\Delta + o_p(\lambda_n)$ and we now show that the optimality conditions are satisfied with rank $\mathbf{r}$. From the regular consistency, the rank of $\hat{W}$ is, with probability tending to one, larger than $\mathbf{r}$ (because the rank is lower semi-continuous function). We now need to show that it is actually equal to $\mathbf{r}$. We let denote $\hat{W} = USV^\top$ the singular value decomposition of $\hat{W}$ and we let denote U_o and V_o the singular vectors corresponding to all but the $\mathbf{r}$ largest singular values. Since we have simultaneous singular value decompositions, we simply need to show that, $\left\| U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon} \right) V_o \right\|_2 < \lambda_n$ with probability tending to one. We have:

$$
\begin{aligned}
U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon} \right) V_o &= U_o^\top \left(\lambda_n \hat{\Sigma}_{mm}\Delta + o_p(\lambda_n) - O_p(n^{-1/2}) \right) V_o \\
&= \lambda_n U_o^\top (\Sigma_{mm}\Delta) V_o + o_p(\lambda_n).
\end{aligned}
$$

Moreover, because of regular consistency and a positive eigengap for $\mathbf{W}$, the projection onto the first $\mathbf{r}$ singular vectors of $\hat{W}$ converges to the projection onto the first $\mathbf{r}$ singular vectors of $\mathbf{W}$ (see Appendix A), which implies that the projection onto the orthogonal is also consistent, that is, $U_o U_o^\top$ converges in probability to $\mathbf{U}_\perp \mathbf{U}_\perp^\top$ and $V_o V_o^\top$ converges in probability to $\mathbf{V}_\perp \mathbf{V}_\perp^\top$. Thus:

$$
\begin{aligned}
\left\| U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon} \right) V_o \right\|_2 &= \left\| U_o U_o^\top \left(\hat{\Sigma}_{mm}(\hat{W} - \mathbf{W}) - \hat{\Sigma}_{M\varepsilon} \right) V_o V_o^\top \right\|_2 \\
&= \lambda_n \| \mathbf{U}_\perp \mathbf{U}_\perp^\top (\Sigma_{mm}\Delta) \mathbf{V}_\perp \mathbf{V}_\perp^\top \|_2 + o_p(\lambda_n) \\
&= \lambda_n \| \Lambda \|_2 + o_p(\lambda_n).
\end{aligned}
$$

This implies that that the last expression is asymptotically of magnitude strictly less than one, which concludes the proof.

C.9 Proof of Theorem 13

We have seen earlier that if $n^{1/2}\lambda_n$ tends to zero and λ_n tends to zero, then Eq. (4) is necessary for rank-consistency. We just have to show that there is a subsequence that does satisfy this. If $\liminf \lambda_n > 0$, then we cannot have consistency (by Proposition 6), thus if we consider a subsequence, we can always assume that λ_n tends to zero.

We now consider the sequence $n^{1/2}\lambda_n$, and its accumulation points. If zero or $+\infty$ is one of them, then by Propositions 7 and 9, we cannot have rank consistency. Thus, for all accumulation points (which are finite and strictly positive), by considering a subsequence, we are in the situation where $n^{1/2}\lambda_n$ tends to $+\infty$ and λ_n tends to zero, which implies Eq. (4), by definition of Λ in Eq. (3) and Lemma 11.

C.10 Proof of Theorem 15

We let denote $U_{LS}^{\mathbf{r}}$ and $V_{LS}^{\mathbf{r}}$ the first $\mathbf{r}$ columns of U_{LS} and V_{LS} and U_{LS}^o and V_{LS}^o the remaining columns; we also denote $s_{LS}^{\mathbf{r}}$ the corresponding first $\mathbf{r}$ singular values and s_{LS}^o the remaining singular values. From Lemma 14 and results in the appendix, we get that $\|s_{LS}^{\mathbf{r}} - \mathbf{s}\|_2 = O_p(n^{-1/2})$ and $\|s_{LS}^o\|_2 = O_p(n^{-1/2})$ and $\|U_{LS}^{\mathbf{r}}(U_{LS}^{\mathbf{r}})^\top - \mathbf{U}\mathbf{U}^\top\|_2 = O_p(n^{-1/2})$ and $\|V_{LS}^{\mathbf{r}}(V_{LS}^{\mathbf{r}})^\top - \mathbf{V}\mathbf{V}^\top\|_2 = O_p(n^{-1/2})$. By writing $\hat{W}_A = \mathbf{W} + n^{-1/2}\hat{\Delta}_A$, $\hat{\Delta}_A$ is defined as the minimum of

$$
\frac{1}{2} \operatorname{vec}(\Delta)^\top \hat{\Sigma}_{mm} \operatorname{vec}(\Delta) - n^{1/2}\operatorname{tr}\Delta^\top \hat{\Sigma}_{M\varepsilon} + n\lambda_n \left(\|A\mathbf{W}B + n^{-1/2}A\Delta B\|_* - \|A\mathbf{W}B\|_* \right).
$$

We have:

$$
\begin{aligned}
A\mathbf{U} &= U_{LS}\mathrm{Diag}(s_{LS})^{-\gamma}U_{LS}^{\top}\mathbf{U} \\
&= U_{LS}^{\mathbf{r}}\mathrm{Diag}(s_{LS}^{\mathbf{r}})^{-\gamma}(U_{LS}^{\mathbf{r}})^{\top}\mathbf{U} + U_{LS}^{o}\mathrm{Diag}(s_{LS}^{o})^{-\gamma}(U_{LS}^{o})^{\top}\mathbf{U} \\
&= \mathbf{U}\mathrm{Diag}(\mathbf{s})^{-\gamma} + O_p(n^{-1/2}) + O_p(n^{-1/2}n^{\gamma/2}) \\
&= \mathbf{U}\mathrm{Diag}(\mathbf{s})^{-\gamma} + O_p(n^{-1/2}n^{\gamma/2}),
\end{aligned}
$$

and

$$
\begin{aligned}
A\mathbf{U}_{\perp} &= U_{LS}\mathrm{Diag}(s_{LS})^{-\gamma}U_{LS}^{\top}\mathbf{U}_{\perp} \\
&= U_{LS}^{\mathbf{r}}\mathrm{Diag}(s_{LS}^{\mathbf{r}})^{-\gamma}(U_{LS}^{\mathbf{r}})^{\top}\mathbf{U}_{\perp} + U_{LS}^{o}\mathrm{Diag}(s_{LS}^{o})^{-\gamma}(U_{LS}^{o})^{\top}\mathbf{U}_{\perp} \\
&= \mathbf{U}_{\perp}\mathrm{Diag}(s_{LS}^{o})^{-\gamma} + O_p(n^{\gamma/2-1/2}) \\
&= O_p(n^{\gamma/2}).
\end{aligned}
$$

Similarly we have: $B\mathbf{V} = \mathbf{V}\mathrm{Diag}(\mathbf{s})^{-\gamma} + O_p(n^{-1/2}n^{\gamma/2})$ and $B\mathbf{V} = O_p(n^{\gamma/2})$. We can decompose any $\Delta \in \mathbb{R}^{p\times q}$ as $\Delta = (\mathbf{U}\ \mathbf{U}_{\perp})\begin{pmatrix}\Delta_{\mathbf{rr}} & \Delta_{\mathbf{r}o} \\ \Delta_{o\mathbf{r}} & \Delta_{oo}\end{pmatrix}(\mathbf{V}\ \mathbf{V}_{\perp})^{\top}$. We have assumed that $\lambda_n n^{1/2}n^{\gamma/2}$ tends to infinity. Thus,

- if $\mathbf{U}_{\perp}^{\top}\Delta = 0$ and $\Delta\mathbf{V}_{\perp} = 0$ (i.e., if Δ is of the form $\mathbf{U}\Delta_{\mathbf{rr}}\mathbf{V}^{\top}$),

$$
\begin{aligned}
n\lambda_n\|A\mathbf{W}B + n^{-1/2}A\Delta B\|_* - \|A\mathbf{W}B\|_* &\leqslant \lambda_n n^{1/2}\|A\Delta B\|_* \\
&= \lambda_n n^{1/2}\|\mathrm{Diag}(\mathbf{s})^{-\gamma}\Delta_{\mathbf{rr}}\mathrm{Diag}(\mathbf{s})^{-\gamma}\|_* \\
&\qquad + O_p(\lambda_n n^{\gamma/2}) \\
&= O_p(\lambda_n n^{1/2})
\end{aligned}
$$

  tends to zero.

- Otherwise, $n\lambda_n\|A\mathbf{W}B + n^{-1/2}A\Delta B\|_* - \|A\mathbf{W}B\|_*$ is larger than $\lambda_n n^{1/2}\|A\Delta B\|_* - 2\|A\mathbf{W}B\|_*$. The term $\|A\mathbf{W}B\|_*$ is bounded in probability because we can write $A\mathbf{W}B = \mathbf{U}\mathrm{Diag}(\mathbf{s})^{1-2\gamma}\mathbf{V}^{\top} + O_p(n^{-1/2+\gamma/2})$ and $\gamma \leqslant 1$. Besides, $\lambda_n n^{1/2}\|A\Delta B\|_*$ is tending to infinity as soons as any of $\Delta_{o\mathbf{r}}, \Delta_{\mathbf{r}o}$ or $\Delta_{\mathbf{rr}}$ are different from zero. Indeed, by equivalence of finite dimensional norms $\lambda_n n^{1/2}\|A\Delta B\|_*$ is larger than a constant times $\lambda_n n^{1/2}\|A\Delta B\|_F$, which can be decomposed in four pieces along $(\mathbf{U},\mathbf{U}_{\perp})$ and $(\mathbf{V},\mathbf{V}_{\perp})$, corresponding asymptotically to $\Delta_{oo}, \Delta_{o\mathbf{r}}, \Delta_{\mathbf{r}o}$ or $\Delta_{\mathbf{rr}}$. The smallest of those terms grows faster than $\lambda_n n^{1/2+\gamma/2}$, and thus tends to infinity.

Thus, since Σ_{mm} is invertible, by the epi-convergence theorem of Geyer (1994, 1996), $\hat{\Delta}_A$ converges in distribution to the minimum of

$$
\frac{1}{2}\mathrm{vec}(\Delta)^{\top}\Sigma_{mm}\mathrm{vec}(\Delta) - n^{1/2}\mathrm{tr}\Delta^{\top}\hat{\Sigma}_{M\varepsilon},
$$

such that $\mathbf{U}_{\perp}^{\top}\Delta = 0$ and $\Delta\mathbf{V}_{\perp} = 0$. This minimum has a simple asymptotic distribution, namely $\Delta = \mathbf{U}\Theta\mathbf{V}^{\top}$ and Θ is asymptotically normal with mean zero and covariance matrix $\sigma^2\left[(\mathbf{V}\otimes\mathbf{U})^{\top}\Sigma_{mm}(\mathbf{V}\otimes\mathbf{U})\right]^{-1}$, which leads to the consistency and the asymptotic normality.

In order to finish the proof, we consider the optimality conditions which can be written as $A\Delta B$ and

$$A^{-1}\left(\hat{\Sigma}_{mm}\hat{\Delta}_A - n^{1/2}\hat{\Sigma}_{M\varepsilon}\right)B^{-1}$$

having simultaneous singular value decompositions with proper decays of singular values, that is, such that the first $\mathbf{r}$ are equal to $\lambda_n n^{1/2}$ and the remaining ones are less than $\lambda_n n^{1/2}$.

From the asymptotic normality we get that $\hat{\Sigma}_{mm}\hat{\Delta}_A - n^{1/2}\hat{\Sigma}_{M\varepsilon}$ is $O_p(1)$, we can thus consider matrices of the form $A^{-1}\Theta B^{-1}$ where Θ is bounded, the same way we considered matrices of the form $A\Delta B$.

We have:

$$
\begin{aligned}
A^{-1}\mathbf{U} &= U_{LS}\mathrm{Diag}(s_{LS})^{\gamma}U_{LS}^{\top}\mathbf{U} \\
&= U_{LS}^{\mathbf{r}}\mathrm{Diag}(s_{LS}^{\mathbf{r}})^{\gamma}(U_{LS}^{\mathbf{r}})^{\top}\mathbf{U} + U_{LS}^{o}\mathrm{Diag}(s_{LS}^{o})^{\gamma}(U_{LS}^{o})^{\top}\mathbf{U} \\
&= \mathbf{U}\mathrm{Diag}(\mathbf{s})^{\gamma} + O_p(n^{-1/2}),
\end{aligned}
$$

and

$$
\begin{aligned}
A^{-1}\mathbf{U}_{\perp} &= U_{LS}\mathrm{Diag}(s_{LS})^{\gamma}U_{LS}^{\top}\mathbf{U}_{\perp} \\
&= U_{LS}^{\mathbf{r}}\mathrm{Diag}(s_{LS}^{\mathbf{r}})^{\gamma}(U_{LS}^{\mathbf{r}})^{\top}\mathbf{U}_{\perp} + U_{LS}^{o}\mathrm{Diag}(s_{LS}^{o})^{\gamma}(U_{LS}^{o})^{\top}\mathbf{U}_{\perp} \\
&= O_p(n^{-1/2}) + \mathbf{U}_{\perp}\mathrm{Diag}(s_{LS}^{o})^{\gamma},
\end{aligned}
$$

with similar expansions for $B^{-1}\mathbf{V}$ and $B^{-1}\mathbf{V}_{\perp}$. We obtain the first order expansion:

$$
\begin{aligned}
A^{-1}\Theta B^{-1} = {}& \mathbf{U}\mathrm{Diag}(\mathbf{s})^{\gamma}\Theta_{\mathbf{rr}}\mathrm{Diag}(\mathbf{s})^{\gamma}\mathbf{V}^{\top} + \mathbf{U}_{\perp}\mathrm{Diag}(s_{LS}^{o})^{\gamma}\Theta_{o\mathbf{r}}\mathrm{Diag}(\mathbf{s})^{\gamma}\mathbf{V}^{\top} \\
& + \mathbf{U}\mathrm{Diag}(\mathbf{s})^{\gamma}\Theta_{\mathbf{r}o}\mathrm{Diag}(s_{LS}^{o})^{\gamma}\mathbf{V}_{\perp}^{\top} + \mathbf{U}_{\perp}\mathrm{Diag}(s_{LS}^{o})^{\gamma}\Theta_{oo}\mathrm{Diag}(s_{LS}^{o})^{\gamma}\mathbf{V}_{\perp}^{\top}
\end{aligned}
$$

Because of the regular consistency, the first term is of the order of $\lambda_n n^{1/2}$ (so that the first $\mathbf{r}$ singular values of $\hat{W}$ are strictly positive), while the three other terms have norms less than $O_p(n^{-\gamma/2})$ which is less than $O_p(n^{1/2}\lambda_n)$ by assumption. This concludes the proof.

References

J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. Low-rank matrix factorization with attributes. Technical Report N24/06/MM, Ecole des Mines de Paris, 2006.

J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. Technical Report HAL-00250231, HAL, 2008.

Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the International Conference on Machine Learning*, 2007.

A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Advances in Neural Information Processing Systems 19*, 2007.

F. R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

F. R. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems 17*, 2004.

J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizbal. *Numerical Optimization Theoretical and Practical Aspects*. Springer, 2003.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Number 3 in CMS Books in Mathematics. Springer-Verlag, 2000.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003.

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

M. Fazel, H. Hindi, and S. P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings American Control Conference*, volume 6, pages 4734–4739, 2001.

W. Fu and K. Knight. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

C. J. Geyer. On the asymptotics of constrained *m*-estimation. *Annals of Statistics*, 22(4):1993–2010, 1994.

C. J. Geyer. On the asymptotics of convex stochastic optimization. Technical report, School of Statistics, University of Minnesota, 1996.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, 1966.

A. S. Lewis and H. S. Sendov. Twice differentiable spectral functions. *SIAM J. Mat. Anal. App.*, 23 (2):368–386, 2002.

Z. Lu, R. Monteiro, and M. Yuan. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Technical report, Optimization online, 2008. URL http://www.optimization-online.org/DB_HTML/2008/01/1877.html.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York, 1998.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. Technical Report 720, Dpt of Statistics, UC Berkeley, 2006.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. Technical Report arXiv:0706.4138v1, arXiv, 2007.

J. D. M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the International Conference on Machine Learning*, 2005.

W. Rudin. *Real and complex analysis, Third edition.* McGraw-Hill, 1987.

J. Shao. *Mathematical Statistics*. Springer, 2003.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal Royal Statististics*, 58(1): 267–288, 1994.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

M. Yuan, A. Ekici, Z. Lu, and R. D. C. Monteiro. Dimension reduction and coefficient estimation in the multivariate linear regression. *Journal of the Royal Statistical Society, Series B*, 69(3): 329–346, 2007.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

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