The Annals of Statistics
2018, Vol. 46, No. 6B, 3481–3509
https://doi.org/10.1214/17-AOS1666

ROBUST LOW-RANK MATRIX ESTIMATION

BY ANDREAS ELSENER AND SARA VAN DE GEER

ETH Zürich

Many results have been proved for various nuclear norm penalized estimators of the uniform sampling matrix completion problem. However, most of these estimators are not robust: in most of the cases the quadratic loss function and its modifications are used. We consider robust nuclear norm penalized estimators using two well-known robust loss functions: the absolute value loss and the Huber loss. Under several conditions on the sparsity of the problem (i.e., the rank of the parameter matrix) and on the regularity of the risk function sharp and nonsharp oracle inequalities for these estimators are shown to hold with high probability. As a consequence, the asymptotic behavior of the estimators is derived. Similar error bounds are obtained under the assumption of weak sparsity, that is, the case where the matrix is assumed to be only approximately low-rank. In all of our results, we consider a high-dimensional setting. In this case, this means that we assume $n \leq pq$. Finally, various simulations confirm our theoretical results.

1. Introduction.

1.1. *Background and motivation.* Netflix, Spotify, Apple Music, Amazon and many other on-line services offer an almost infinite amount of songs or films to their users. Clearly, a single person will never be able to watch every film or to listen to every available song. For this reason, an elaborate recommendation system is necessary in order to allow the users to choose content that already match his or her preferences. Many models and estimation methods have been proposed to address this question. Matrices provide an appropriate way of modelling this problem. Imagine that the plethora of films/songs is identified with the rows of a matrix, call it B^*, and the users with its columns. One entry of the matrix corresponds to the rating given to film "i" (row) by user "j" (column). This matrix will have many missing entries. These entries are bounded and we can expect the rows of B^* to be very similar to each other. It is therefore sensible to assume that B^* has a low rank. The challenge is now to predict the missing ratings/fill in the empty entries of B^*. Define for this purpose the set of observed (possibly noisy) entries

$$\begin{aligned}\mathcal{A} := \big\{&(i, j) \in \{1, \ldots, p\} \times \{1, \ldots, q\} : \text{the (noisy) entry} \\ &A_{ij} \text{ of } B^* \text{ is observed}\big\},\end{aligned} \tag{1.1}$$

Received May 2016; revised October 2017.
MSC2010 subject classifications. Primary 62J05, 62F30; secondary 62H12.
Key words and phrases. Matrix completion, robustness, empirical risk minimization, oracle inequality, nuclear norm, sparsity.

where p is the number of films/songs and q the number of users. Our estimation problem can therefore be stated in the following way: for $B \in \mathcal{B} \subset \mathbb{R}^{p\times q}$ we

$$\text{minimize } R_n(B), \text{ subject to } \operatorname{rank}(B) = s.$$

In this special case, we have that

$$\mathcal{B} = \{B \in \mathbb{R}^{p\times q} | \|B\|_\infty \le \eta\}, \tag{1.2}$$

where η is for instance the mean highest rating, and R_n is some convex empirical error measure that is defined by the data, for example,

$$R_n(B) = \frac{1}{|\mathcal{A}|} \sum_{(i,j)\in\mathcal{A}} (A_{ij} - B_{ij})^2. \tag{1.3}$$

Since the rank of a matrix is not convex, we use the nuclear norm as its convex surrogate. This leads us to a relaxed convex optimization problem. For $B \in \mathcal{B}$, we

$$\text{minimize } R_n(B), \text{ subject to } \|B\|_{\text{nuclear}} \le \tau,$$

for some $\tau > 0$. The model described above can be considered as a special case of the trace regression model.

In the *trace regression model* [see, e.g., Rohde and Tsybakov (2011)], one considers the observations (X_i, Y_i) satisfying

$$Y_i = \operatorname{trace}(X_i B^*) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{1.4}$$

where ε_i are i.i.d. random errors. The matrices X_i are so-called masks. They are assumed to lie in

$$\chi = \{e_k(q)e_l(p)^T : 1 \le k \le q, 1 \le l \le p\}, \tag{1.5}$$

where $e_k(q)$ is the q-dimensional kth unit vector and $e_l(p)$ is the p-dimensional lth unit vector. We will assume that the X_i are i.i.d. with

$$\mathbb{P}(X_{i_{kj}} = 1) = 1 - \mathbb{P}(X_{i_{kj}} = 0) = \frac{1}{pq}$$

for all $i \in \{1, \dots, n\}$, $k \in \{1, \dots, q\}$, and $j \in \{1, \dots, p\}$. However, we point out that it is not necessary for our estimators to know this distribution. This knowledge will only be used in the proofs of the theoretical results. In Klopp (2014) the case of more general sampling distributions under quadratic loss is treated.

The trace regression model together with the space χ and the distribution on χ is equivalent to the matrix completion case. The entries of the vector Y can be identified with the observed entries as those in the matrix A.

From this, it can be seen that we are in a high-dimensional setting since the number of observations n must be smaller than or equal to the total number of entries of A. The setup described above is then called *uniform sampling matrix completion*. A very similar setup was first considered in Srebro, Rennie and Jaakkola (2004) and Srebro and Shraibman (2005).

As in the standard regression setting, parameter estimation in the trace regression model is also done via empirical risk minimization. Using the Lagrangian form for $B \in \mathcal{B}$, we

$$\text{minimize } R_n(B) + \lambda \|B\|_{\text{nuclear}},$$

where $R_n(B) = 1/n \sum_{i=1}^n \rho(Y_i - \text{trace}(X_i B))$, ρ is a convex loss function and $\lambda > 0$ is the tuning parameter. The loss function is often chosen to be the quadratic loss (or one of its modifications) as in Koltchinskii, Lounici and Tsybakov (2011), Negahban and Wainwright (2011, 2012), Rohde and Tsybakov (2011) and many others. In Lafond (2015), the case of an error distribution belonging to an exponential family is considered. As long as the errors are assumed to be light tailed as it is the case for i.i.d. Gaussian errors the least squares estimator will perform very well. However, the ratings are heavily subject to frauds (e.g., by the producer of a film). It is necessary to take this fact into account also in the estimation procedure. One might also be interested in estimating the median or another quantile of the ratings. For this purpose, M-estimators based on different losses than the quadratic loss are usually chosen.

1.2. *Proposed estimators.* In this paper, we consider the absolute value loss and the Huber loss. The first robust estimator is then given by

$$\hat{B} := \arg\min_{B \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n |Y_i - \text{trace}(X_i B)| + \lambda \|B\|_{\text{nuclear}}. \tag{1.6}$$

Using the Huber loss, we can define

$$\hat{B}_H := \arg\min_{B \in \mathcal{B}} \frac{1}{n} \sum_{i=1}^n \rho_H\big(Y_i - \text{trace}(X_i B)\big) + \lambda \|B\|_{\text{nuclear}}, \tag{1.7}$$

where the function

$$\rho_H(x) := \begin{cases} x^2 & \text{if } |x| \le \kappa, \\ 2\kappa|x| - \kappa^2 & \text{if } |x| > \kappa, \end{cases}$$

defines the Huber loss function. The tuning parameter $\kappa > 0$ is assumed to be given for our estimation problem. The possible values for the Huber parameter κ depend on the distribution of the errors as shown in Lemma 2.1. In practice, one usually estimates κ and λ with methods such as cross-validation. Notice that it could happen that the estimators defined in equations (1.6) and (1.7) are not unique since the objective functions are not strictly convex. As will be shown, the rates depend on the Lipschitz constants of the loss functions and on η. Typically, the Lipschitz constants of the absolute value loss as well as of the Huber loss induce smaller constants in the rates compared to the Lipschitz constant of the truncated quadratic loss.

The "target" is defined as

$$B^0 := \arg\min_{B \in \mathcal{B}} R(B),$$

where $R(B) = \mathbb{E}R_n(B)$ is the theoretical risk. It has to be noticed that the matrix B^0 is not necessarily equal to the matrix B^*. Our main interest lies in the theoretical analysis of the above estimators. The estimators should mimic the sparsity behavior of an oracle B. In the case of the absolute value loss, we will prove a non-sharp oracle inequality. Whereas for the Huber loss, thanks to its differentiability, we are able to derive a sharp oracle inequality.

Assuming $B = B^0$ in Corollary 3.1, the upper bound is typically of the form

$$R(\hat{B}_H) - R(B^0) \lesssim \lambda^2 pqs_0, \tag{1.8}$$

where $\lesssim$ means that some multiplicative constants (depending on the tuning parameter κ) are omitted.

The assumptions for this kind of results are mainly based on the regularity of the absolute value and the Huber loss. Moreover, the properties of the nuclear norm, which are very similar to those of the ℓ_1-norm for vectors, will be exploited. In addition, we use the sparsity behavior induced by the nuclear norm to infer the behavior of weakly sparse estimators. This takes into account that a matrix could have few very large singular values and many small, but not exactly zero singular values:

$$B \in \left\{ B' \in \mathcal{B} : \sum_{j=1}^{q} \Lambda_j^r \le \rho_r^r, \Lambda_1, \ldots, \Lambda_q \text{ the singular values of } B' \right\},$$

where $0 < r < 1$ and ρ_r^r is some reasonably small constant.

1.3. *Related literature.* A first study with robust matrix estimation was made in Chandrasekaran et al. (2011) in a setting with no missing entries. In order to avoid identifiability issues, the authors introduce "incoherence" conditions on the low-rank component. These conditions make sure that the low-rank component itself is not too sparse. The locations of the corruptions are assumed to be fixed. In the context of Principal Component Analysis (PCA) which is a special case of the matrix regression model, robustness was investigated in Candès et al. (2011). The authors assume that the matrix to be estimated is decomposed in a low-rank matrix and a sparse matrix. In contrast to Chandrasekaran et al. (2011), the nonzero entries of the sparse matrix are assumed to be drawn randomly following a uniform distribution. Following this line of research, Li (2013) apply these conditions to the matrix completion problem with randomly observed entries. In a parallel work, Chen et al. (2013) consider the case where the indices of the observed entries may be simultaneously both random and deterministic. In these papers, only noiseless robust matrix completion is considered.

Cambier and Absil (2016) study computational aspects of robust matrix completion (in the previously mentioned setting). A method relying on Riemannian optimization is proposed. The authors assume the rank of the matrix to be estimated to be known.

The robust matrix completion problem in a low-rank plus sparse framework is considered from a computational point of view also in Cherapanamjeri, Gupta and Jain (2016). The authors propose an algorithm based on projected gradient descent and apply it directly to the nonconvex optimization problem. Also in this work, the corruptions are assumed to be limited to the sparse component.

In Foygel et al. (2011) weighted nuclear norm penalized estimators with (possibly nonconvex) Lipschitz continuous loss functions are studied from a learning point of view. The partially observed entries are assumed to follow a possibly nonuniform distribution on the set χ. In contrast, our derivations rely among other properties on the convexity of the risk (i.e., the margin conditions).

Noisy robust matrix completion was first investigated in Klopp, Lounici and Tsybakov (2016). The authors assume that the truth A^* is decomposed in a low-rank matrix and a sparse matrix where the low-rank matrix contains the "parameters of interest" and the sparse matrix contains the corruptions. In addition, every observation is corrupted by independent and centered sub-Gaussian noise. The largest entries of both the low-rank and sparse matrices are assumed to be bounded (e.g., by the maximal possible rating). Their model is as follows: $(X_i, \tilde{Y}_i), i = 1, \dots, N$ satisfy

$$\tilde{Y}_i = \operatorname{trace}\bigl(X_i A^*\bigr) + \xi_i, \qquad i = 1, \dots, N, \tag{1.9}$$

where $A^* = L^* + S^*$ with L^* a low-rank matrix and S^* a matrix with entrywise sparsity. Columnwise sparsity is also considered but in view of a comparison of this and our approach we prefer to restrict to entrywise sparsity. The masks X_i are assumed to lie in the set χ (1.5) and to be independent of the noise ξ_i for all i. The set of observed indices is assumed to be the union of two disjoint components Ξ and $\tilde{\Xi}$. The set Ξ corresponds to the noncorrupted noisy observations (i.e., only entries of L^* plus ξ_i). The entries corresponding to these observations of S^* are zero. The set $\tilde{\Xi}$ contains the indices of the observations that are corrupted by a (nonzero) entry of S^*. It is not known if an observation comes from Ξ or $\tilde{\Xi}$. The estimator given in Klopp, Lounici and Tsybakov (2016) is

$$\begin{aligned}(\hat{L}, \hat{S}) \in \operatorname*{arg\,min}_{\substack{\|L\|_\infty \le \eta \\ \|S\|_\infty \le \eta}} \Biggl\{ & \frac{1}{N} \sum_{i=1}^{N} \bigl(\tilde{Y}_i - \operatorname{trace}\bigl(X_i (L + S)\bigr)\bigr)^2 \\ & + \lambda_1 \|L\|_{\text{nuclear}} + \lambda_2 \|S\|_1 \Biggr\},\end{aligned} \tag{1.10}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters.

In contrast to the previously mentioned papers on robust matrix completion, we consider (possibly heavy-tailed) random errors that affect the observations but not the truth.

1.4. *Organization of the paper.* The paper is organized in the following way. We state the main assumptions that are used throughout the paper in Section 2. Then the nuclear norm, its properties and its similarities to the ℓ_1-norm are discussed. To bound the empirical process part resulting from the matrix completion setting, we make use of the results in Section 3 of the Supplementary Material [Elsener and van de Geer (2018)]. In Section 3, the main theorems are presented: the (deterministic) sharp and nonsharp oracle inequalities. In Section 3.3, we present the applications of these results to the case of Huber loss and absolute value loss. The asymptotics and the applications to weak sparsity are presented in Section 4. Finally, to verify the theoretical findings, Section 5 presents some simulations. The Student t distribution with three degrees of freedom is considered as an error distribution.

2. Preliminaries. In this section, the assumptions on the loss functions, the risk and the distribution of the errors are presented. In particular, Assumptions 2 and 3 below are on the curvature of the (theoretical) risk. They are used to derive the *deterministic* sharp and nonsharp oracle inequalities. It is important to notice that the curvature of the risk mainly depends on the properties of the distribution of the errors. Assumptions 4 and 5 below will be shown to be sufficient for Assumptions 2 and 3 to hold, respectively.

Furthermore, we also discuss the properties of the nuclear norm. Thanks to the penalization term in the objective functions the optimization problems become computationally tractable. We also highlight the commonalities of the vector ℓ_1-norm and the nuclear norm for matrices.

2.1. *Assumptions on the risk and the distribution of the errors.* The first assumption is about the loss function.

ASSUMPTION 1. Let ρ be the loss function. We assume that it is Lipschitz continuous with constant L, that is, that for all $x, y \in \mathbb{R}$,

$$|\rho(x) - \rho(y)| \leq L|x - y|. \tag{2.1}$$

The next two assumptions ensure the identifiability of the parameters by requiring a sufficient convexity of the risk around the target.

ASSUMPTION 2. One-point-margin condition. There is an increasing strictly convex function G with $G(0) = 0$ such that for all $B \in \mathcal{B}$

$$R(B) - R(B^0) \geq G(\|B - B^0\|_F),$$

where R is the theoretical risk function.

ASSUMPTION 3. Two-point-margin condition. There is an increasing strictly convex function G with $G(0) = 0$, such that for all $B, B' \in \mathcal{B}$, we have

$$R(B) - R(B') \geq \operatorname{trace}\big(\dot{R}(B')^T (B - B')\big) + G\big(\|B - B'\|_F\big),$$

where R is the theoretical risk function and $[\dot{R}(B')]_{kl} = \frac{\partial}{\partial B_{kl}} R(B)|_{B=B'}$.

Assumption 1 is crucial when it comes to the application of the contraction theorem which in turn allows us to apply the dual norm inequality to find a bound for the random part of the oracle bounds. Assumptions 2 and 3 are essential in the proofs of the (deterministic) results. In particular, in addition to the differentiability of the empirical risk R_n, Assumption 3 is responsible for the sharpness of the first oracle bound that will be proved. The margin conditions are strongly related to the shape of the distribution function and the corresponding density of the errors.

For the specific application to the Huber loss and absolute value loss estimators, we show that mild conditions on the distribution of the errors ensure a sufficient curvature of the risk for both loss functions under study.

Assumption 3 holds under a weak condition on the distribution function of the errors.

ASSUMPTION 4. Assume that there exists a constant $C_1 > 0$ such that the distribution function F with density with respect to Lebesgue measure f of the errors fulfills

$$F(u + \kappa) - F(u - \kappa) \geq 1/C_1^2, \qquad \text{for all } |u| \leq 2\eta \text{ and } \kappa \leq 2\eta. \tag{2.2}$$

LEMMA 2.1. *Assumption 4 implies Assumption 3 with* $G(u) = u^2/(2C_1^2 pq)$.

The following assumption guarantees that Assumption 2 holds.

ASSUMPTION 5. Suppose $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with median zero and density f with respect to Lebesgue measure. Assume that for $C_2 > 0$

$$f(u) \geq \frac{1}{C_2^2}, \qquad \text{for all } |u| \leq 2\eta. \tag{2.3}$$

LEMMA 2.2. *Assumption 5 implies Assumption 2 with* $G(u) = u^2/(2C_2^2 pq)$.

Another important fact is that when the distribution of the errors is assumed to be symmetric around zero $B^* = B^0$. This phenomenon is discussed in Section 4 of the Supplementary Material [Elsener and van de Geer (2018)].

2.2. *Properties of the nuclear norm.* The regularization by the nuclear norm plays a similar role as the ℓ_1-norm in the Lasso [Tibshirani (1996)]. We illustrate the similarities and differences of these types of regularizations. In view of the oracle inequalities and in order to keep the notation as simple as possible, we discuss the properties of the nuclear norm of the oracle. The oracle is typically a value B that takes an up-to-constants optimal trade-off between approximation error and estimation error. In what follows, B is called "the oracle" although its choice is flexible.

Consider the singular value decomposition of the oracle B with rank $s^\star$ given by

$$B = P\Lambda Q^T, \tag{2.4}$$

where P is a $p \times q$ matrix, Q a $q \times p$ matrix and Λ a $q \times q$ diagonal matrix containing the ordered singular values $\Lambda_1 \geq \cdots \geq \Lambda_q$. Then the nuclear norm is given by

$$\|B\|_{\text{nuclear}} = \sum_{i=1}^{q} \Lambda_i(B) = \big\|\Lambda(B)\big\|_1, \tag{2.5}$$

by interpreting $\Lambda(B) \in \mathbb{R}^q$ as the vector of singular values. The penalization with the nuclear norm induces sparsity in the singular values, whereas the penalization with the vector ℓ_1-norm of the parameters in linear regression induces sparsity directly in the parameters. On the other hand, the rank plays the role of the number of nonzero coefficients in the Lasso setting, namely

$$s^\star = \big\|\Lambda(B)\big\|_0. \tag{2.6}$$

One main ingredient of the proofs of the oracle inequalities is the so-called triangle property as introduced in van de Geer (2001). This property was used in, for example, Bühlmann and van de Geer (2011) to prove nonsharp oracle inequalities. For the ℓ_1-norm, the triangle property follows from its decomposability. For the nuclear norm, the triangle property as it is used in this work depends on the features of the oracle B. For this reason, we notice that for any positive integer $s \leq q$ the oracle can be decomposed in

$$B = B^+ + B^-, \qquad B^+ = \sum_{k=1}^{s} \Lambda_k P_k Q_k^T, \qquad B^- = \sum_{k=s+1}^{q} \Lambda_k P_k Q_k^T.$$

The matrix B^+ is called "active" part of the oracle B, whereas the matrix B^- is called the "nonactive" part. The singular value decomposition of B^+ is given by

$$B^+ = P^+ \Lambda Q^{+^T}.$$

We observe that the integer s is not necessarily the rank of the oracle B. The choice of s is free. One may choose a value that trades off the roles of the "active" part B^+ and "nonactive" part B^-; see Lemma 4.1. The following lemma is adapted from Lemma 7.2 and Lemma 12.5 in van de Geer (2016).

LEMMA 2.3. *Let $B^+ \in \mathbb{R}^{p\times q}$ be the active part of the oracle B. Then we have for all $B' \in \mathbb{R}^{p\times q}$ with*

$$\Omega^+_{B^+}(B') := \sqrt{s}\big(\|P^+P^{+^T}B'\|_F + \|B'Q^+Q^{+^T}\|_F + \|P^+P^{+^T}B'Q^+Q^{+^T}\|_F\big)$$

and

$$\Omega^-_{B^+}(B') := \|(I - P^+P^{+^T})B'(I - Q^+Q^{+^T})\|_{\text{nuclear}}$$

that

$$\|B^+\|_{\text{nuclear}} - \|B'\|_{\text{nuclear}} \le \Omega^+_{B^+}(B' - B^+) - \Omega^-_{B^+}(B'). \tag{2.7}$$

We then say that the triangle property holds at B^+.

In particular, since $\Omega^+_{B^+}(B^-) = 0$, we have for any $B' \in \mathbb{R}^{p\times q}$

$$\|B\|_{\text{nuclear}} - \|B'\|_{\text{nuclear}} \le \Omega^+_{B^+}(B' - B) - \Omega^-_{B^+}(B' - B) + 2\|B^-\|_{\text{nuclear}}. \tag{2.8}$$

Moreover, we have

$$\|\cdot\|_{\text{nuclear}} \le \Omega^+_{B^+} + \Omega^-_{B^+}. \tag{2.9}$$

From now on, we write $\Omega^+ = \Omega^+_{B^+}$ and $\Omega^- = \Omega^-_{B^+}$. Equation (2.8) is proved in Appendix A.

Hence, the property that our estimators should mimic is not the rank of the oracle but rather the fact that the "nonactive" part is zero under the semi-norm induced by the active part.

Moreover, we define the norm $\underline{\Omega}$ as

$$\underline{\Omega} := \Omega^+ + \Omega^-.$$

REMARK 1. Notice that the semi-norms Ω^+ and Ω^- form a complete pair, meaning that $\underline{\Omega} := \Omega^+ + \Omega^-$ is a norm.

The estimation error in several different norms can thus be "computed" in general (semi-)norms.

A tail bound for the maximal singular value of a finite sum of matrices lying in the set χ defined in equation (1.5) is given in the following theorem. For this purpose, we first need to define the Orlicz norm of a random variable. Let $Z \in \mathbb{R}$ be a random variable and $\alpha \ge 1$ a constant. Then the Ψ_α-Orlicz norm is defined as

$$\|Z\|_{\Psi_\alpha} := \inf\{c > 0 : \mathbb{E}\exp[|Z|^\alpha/c^\alpha] \le 2\}. \tag{2.10}$$

THEOREM 2.1 [Proposition 2 in Koltchinskii, Lounici and Tsybakov (2011)]. *Let $X_1, \dots, X_n$ be i.i.d. $q \times p$ matrices that satisfy for some $\alpha \ge 1$ (and all i)*

$$\mathbb{E}X_i = 0, \qquad \|\Lambda_{\max}(X_i)\|_{\Psi_\alpha} =: K < \infty.$$

Define

$$S^2 := \max\left\{\Lambda_{\max}\left(\sum_{i=1}^n \mathbb{E}X_i X_i^T\right)\Big/n, \Lambda_{\max}\left(\sum_{i=1}^n \mathbb{E}X_i^T X_i\right)\Big/n\right\}.$$

Then for a constant C and for all $t > 0$,

$$\mathbb{P}\left(\Lambda_{\max}\left(\sum_{i=1}^n X_i\right)\Big/n \geq CS\sqrt{\frac{t+\log(p+q)}{n}} + C\log^{1/\alpha}\left(\frac{K}{S}\right)\left(\frac{t+\log(p+q)}{n}\right)\right) \leq \exp(-t).$$

This theorem is used with the tail summation property of the expectation in the derivations of the tail bounds in Section 3 of the Supplementary Material [Elsener and van de Geer (2018)].

3. Oracle inequalities. We first give two deterministic sharp and nonsharp oracle inequalities. The connection to the empirical process parts and to the specific loss functions follow in Section 3.3. Let $B^0 = \arg\min_{B' \in \mathcal{B}} R(B')$ be the target. It is assumed that $q \leq p$.

3.1. *Sharp oracle inequality.* Here, we assume that the loss function is differentiable and Lipschitz continuous. The next lemma gives a connection between the empirical risk and the penalization term.

LEMMA 3.1 [Adapted from Lemma 7.1 in van de Geer (2016)]. *Suppose that R_n is differentiable. Then for all $B \in \mathcal{B}$*

$$-\operatorname{trace}\big(\dot{R}_n(\hat{B})^T(B-\hat{B})\big) \leq \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}}.$$

The following theorem is inspired by Theorem 7.1 in van de Geer (2016). In contrast to this theorem, we need to bound the empirical process part differently. In view of the application to the matrix completion problem, we assume a specific bound on the empirical process.

THEOREM 3.1. *Suppose that Assumptions 1 and 3 hold, that the loss function is differentiable and let H be the convex conjugate of G. Assume further for all $B' \in \mathcal{B}$ that for $\lambda_\varepsilon > 0$ and $\lambda_* > 0$*

$$\big|\operatorname{trace}\big((\dot{R}_n(B') - \dot{R}(B'))^T(B-B')\big)\big| \leq \lambda_\varepsilon \underline{\Omega}(B'-B) + \lambda_*.$$

Take $\lambda > \lambda_\varepsilon$. Let $0 \leq \delta < 1$ be arbitrary, and define

$$\underline{\lambda} := \lambda - \lambda_\varepsilon, \qquad \overline{\lambda} := \lambda_\varepsilon + \lambda + \delta\underline{\lambda}.$$

Then

$$\delta\underline{\lambda}\Omega^{+}(\hat{B}-B)+\delta\underline{\lambda}\Omega^{-}(\hat{B}-B)+R(\hat{B})-R(B)$$
$$\leq H(\overline{\lambda}3\sqrt{s})+2\lambda\|B^{-}\|_{\text{nuclear}}+\lambda_{*}.$$

In the proof of this theorem, the differentiability of the loss function and Assumption 3 are crucial. Without this property, an additional term arising from the one-point-margin condition would appear in the upper bound. This term would then lead to a nonsharp bound.

3.2. *Nonsharp oracle inequality.* Instead of bounding an empirical process term depending on the derivative of the empirical and theoretical risks, we need to consider differences of these functions.

THEOREM 3.2. *Suppose that Assumptions 1 and 2 hold. Let H be the convex conjugate of G. Suppose further that for $\lambda_{\varepsilon}>0$, $\lambda_{*}>0$, and all $B'\in\mathcal{B}$*

$$\big|\big[R_n(B')-R(B')\big]-\big[R_n(B)-R(B)\big]\big|\leq\lambda_{\varepsilon}\underline{\Omega}(B'-B)+\lambda_{*}. \tag{3.1}$$

Let $0<\delta<1$, take $\lambda>\lambda_{\varepsilon}$ and define

$$\overline{\lambda}=\lambda+\lambda_{\varepsilon},\qquad \underline{\lambda}=\lambda-\lambda_{\varepsilon}. \tag{3.2}$$

Then

$$\delta\underline{\lambda}\underline{\Omega}(\hat{B}-B)\leq 2H\big(\overline{\lambda}(1+\delta)3\sqrt{s}\big)$$
$$+2\big(\lambda_{*}+\big(R(B)-R(B^{0})\big)\big)+4\lambda\|B^{-}\|_{\text{nuclear}}$$

and

$$R(\hat{B})-R(B)\leq\frac{1}{\delta}\big[2H\big(\overline{\lambda}(1+\delta)3\sqrt{s}\big)+\lambda_{*}+2\big(R(B)-R(B^{0})\big)$$
$$+2\lambda\|B^{-}\|_{\text{nuclear}}\big]+\lambda_{*}+2\lambda\|B^{-}\|_{\text{nuclear}}.$$

It has to be noticed that the above bound is "good" only if $R(B)-R(B^{0})$ is already small. The main cause for the nonsharpness is Assumption 2 that leads to an additional term in the upper bound of the inequality.

3.3. *Applications to specific loss functions.* We now apply the deterministic sharp and nonsharp oracle inequalities to the case of the Huber loss and absolute value loss, respectively. We assume in both cases that the distribution of the errors is symmetric around 0 so that $B^{0}=B^{*}$. This is discussed in detail in Section 4 of the Supplementary Material [Elsener and van de Geer (2018)].

Huber loss-sharp oracle inequality. We first consider the case that arises by choosing the Huber loss. Theorem 3.1 together with Lemma 2.1 and the first claim of Lemma 3.2 in the Supplementary Material imply the following corollary. It is useful to notice that the Lipschitz constant of the Huber loss is 2κ.

COROLLARY 3.1. *Let $B = B^+ + B^-$ where B^+ and B^- are defined in equation (2.4). Let Assumption 4 be satisfied.*

For a constant $C_0 > 0$, let

$$\lambda_\varepsilon = 2(4\eta + 2\kappa)\left((8C_0 + \sqrt{2})\sqrt{\frac{\log(p+q)}{nq}} + 8C_0\sqrt{\log(1+q)}\frac{\log(p+q)}{n}\right)$$

and $\lambda_* = 8\eta(4\eta + 2\kappa)p\log(p+q)/(3n) + \lambda_\varepsilon\sqrt{\log(p+q)/n}$.

Assume that $\lambda > \lambda_\varepsilon$. Take $0 \le \delta < 1$,

$$\underline{\lambda} := \lambda - \lambda_\varepsilon \quad \text{and} \quad \overline{\lambda} := \lambda_\varepsilon + \lambda + \delta\underline{\lambda} \tag{3.3}$$

Choose $j_0 := \lceil \log_2(7q\sqrt{pq}\eta\sqrt{n/\log(p+q)}) \rceil$ and define

$$\alpha = (j_0 + 2)\exp\bigl(-p\log(p+q)\bigr).$$

Then we have with probability at least $1 - \alpha$ that

$$\delta\underline{\lambda}\Omega^+(\hat{B}_H - B) + \delta\underline{\lambda}\Omega^-(\hat{B}_H - B) + R(\hat{B}_H) - R(B)$$
$$\le \frac{pqC_1^2\bar{\lambda}^2 9s}{2} + 2\lambda\|B^-\|_{\text{nuclear}} + \lambda_*.$$

Assumption 4 guarantees that the risk function is sufficiently convex. From this assumption, we also obtain a bound for the possible values of the tuning parameter κ. We can also see that the results hold for errors with a heavier tail than the Gaussian. The choice of the noise level λ_ε and consequently of the tuning parameter λ results from the the probability inequalities for the empirical process in Section 3 of the Supplementary Material [Elsener and van de Geer (2018)]. The quantity λ_* is also a consequence of the bound on the empirical process part. However, it does not affect the asymptotic rates.

Absolute value loss—nonsharp oracle inequality. The next corollary is an application to the case of the absolute value loss. Theorem 3.2 combined with Lemma 2.2 and the second claim of Lemma 3.2 in the Supplementary Material lead to the following corollary. The Lipschitz constant in this case is 1.

COROLLARY 3.2. *Let the oracle B be as in 2.4. Suppose that Assumption 5 is satisfied. For a constant $C_0 > 0$, let*

$$\lambda_\varepsilon = 2\left((8C_0 + \sqrt{2})\sqrt{\frac{\log(p+q)}{nq}} + 8C_0\sqrt{\log(1+q)}\frac{\log(p+q)}{n}\right)$$

and $\lambda_* = 8\eta p \log(p+q)/(3n) + \lambda_\varepsilon \sqrt{\log(p+q)/n}$. *Take* $0 < \delta < 1$ *and* $\lambda > \lambda_\varepsilon$. *Choose* $j_0 := \lceil \log_2(7q\sqrt{pq}\eta\sqrt{n/\log(p+q)}) \rceil$ *and define*

$$\alpha = (j_0 + 2)\exp\bigl(-p\log(p+q)\bigr).$$

Then we have with probability at least $1 - \alpha$ *that*

$$\begin{aligned}&\delta\underline{\lambda\Omega}(\hat{B} - B)\\ &\quad\le 6C^2\overline{\lambda}^2(1+\delta)^2 pqs + 2\lambda_* + 2\bigl(R(B) - R\bigl(B^0\bigr)\bigr) + 4\lambda\bigl\|B^-\bigr\|_{\text{nuclear}}\end{aligned}$$

and

$$\begin{aligned}R(\hat{B}) - R(B) \le \frac{1}{\delta}\bigl[&6C_2^2\overline{\lambda}^2(1+\delta)^2 pqs + \lambda_* + 2\bigl(R(B) - R\bigl(B^0\bigr)\bigr)\\ &+ 2\lambda\bigl\|B^-\bigr\|_{\text{nuclear}}\bigr] + \lambda_* + 2\lambda\bigl\|B^-\bigr\|_{\text{nuclear}}.\end{aligned}$$

Also in this case, the choices of λ_ε and λ_* are a consequence of the probability bounds.

4. Asymptotics and weak sparsity. The results in Section 3 are valid for finite values of the dimension of the matrix p, q, the rank and the number of observed entries n. A question that is answered in this section is how the estimation errors of the proposed estimators behave when n, p and q are allowed to grow.

As mentioned in Negahban and Wainwright (2012), practical reasons motivate the assumption that the matrix B^0 is not exactly low-rank but only approximately. In relation to the matrix completion problem, one observes that the ratings given by the users are unlikely to be exactly equal but rather very similar. This translates to a matrix that is not low-rank. However, it is sensible to assume that the matrix is almost low-rank. The notion of weak sparsity quantifies this assumption by assuming that for some $0 < r < 1$ and $\rho > 0$,

$$\sum_{k=1}^{q} \bigl(\Lambda_k^0\bigr)^r =: \rho_r^r, \tag{4.1}$$

where $\Lambda_1^0, \ldots, \Lambda_q^0$ are the singular values of B^0. For $r = 0$, we have under the convention that $0^0 = 0$ that

$$\sum_{k=1}^{q} \bigl(\Lambda_k^0\bigr)^0 = \sum_{k=1}^{q} \mathbb{1}_{\{\Lambda_k^0 > 0\}} = s_0,$$

where s_0 is the rank of B^0. The following lemma gives a bound of the nonactive part of the matrix B that appears in the oracle bounds.

LEMMA 4.1. *For $\sigma > 0$, we may take*

$$\|B^-\|_{\text{nuclear}} \leq \sigma^{1-r}\rho_r^r, \tag{4.2}$$

and

$$s \leq \sigma^{-r}\rho_r^r.$$

We first consider the asymptotic behavior of our estimators in the case of an exactly low-rank matrix and deduce from this the asymptotics for the case of an approximately low-rank matrix.

4.1. *Asymptotics.*

4.1.1. *Sharp.* By Corollary 3.1, assuming that $q\log(1+q) = o(\frac{n}{\log(p+q)})$ and, therefore, using the choice for the noise level

$$\lambda_\varepsilon \asymp \sqrt{\frac{\log(p+q)}{nq}}$$

we obtain

$$\begin{aligned} &R(\hat{B}_H) - R(B^0) \\ &\quad \leq R(B) - R(B^0) \\ &\qquad + \mathcal{O}_{\mathbb{P}}\Bigg(\frac{ps\log(p+q)}{n} \\ &\qquad + \sqrt{\frac{\log(p+q)}{nq}}\Bigg(\sqrt{\frac{\log(p+q)}{n}} + \|B^-\|_{\text{nuclear}}\Bigg)\Bigg). \end{aligned} \tag{4.3}$$

We choose for simplicity the oracle to be the matrix B^0 itself with $s_0 = \operatorname{rank}(B^0)$. Then we make use of the two point margin condition that is shown to hold in Lemma 2.1. The resulting rate is then given by

$$\|\hat{B}_H - B^0\|_F^2 = \mathcal{O}_{\mathbb{P}}\Bigg((4\eta + 2\kappa)^2 C_1^4 \frac{p^2 q s_0 \log(p+q)}{n}\Bigg), \tag{4.4}$$

where κ is the Huber parameter and C_1 is the constant from Lemma 2.1.

REMARK 2. The rate (4.4) depends on η as in Koltchinskii, Lounici and Tsybakov (2011) and on the Lipschitz constant of the loss function which is typically smaller than η. If $C_1^2 = O(\eta)$, the constant in front of the rate is of order $O(\eta^4)$. This is a "worst-case" scenario that shows the cost that is paid when allowing for very general error distributions as in our case. We emphasize that in this case the distribution of the errors is not required to have a density.

In addition to the rate obtained for the Frobenius norm, we are also able to derive rates for the estimation error measured in nuclear norm. From Corollary 3.1 and equation (2.9) under the previous conditions, it follows that

$$\|\hat{B}_H - B^0\|_{\text{nuclear}} = \mathcal{O}_{\mathbb{P}}\left(C_1^2 pqs_0\sqrt{\frac{\log(p+q)}{nq}}\right). \tag{4.5}$$

4.1.2. *Nonsharp*. By Corollary 3.2, it is known that the assumption $q\log(1+q) = o(\frac{n}{\log(p+q)})$ leads to the choice $\lambda_\varepsilon \asymp \sqrt{\log(p+q)/nq}$. Therefore, we

$$\begin{aligned} &R(\hat{B}) - R\big(B^0\big) \\ &\quad = \mathcal{O}_{\mathbb{P}}\left(\frac{ps\log(p+q)}{n} + R(B) - R\big(B^0\big)\right. \\ &\qquad \left. + \sqrt{\frac{\log(p+q)}{nq}}\left(\sqrt{\frac{\log(p+q)}{n}} + \|B^-\|_{\text{nuclear}}\right)\right). \end{aligned} \tag{4.6}$$

What can be observed comparing the rates in equations (4.3) and (4.6) is the presence of the additional term $R(B) - R(B^0)$ in the nonsharp case in contrast to the sharp case. We choose again the oracle to be the matrix B^0 itself. By the one point margin condition derived in Lemma 2.2, we see that the rate of convergence in this case is given by

$$\|\hat{B} - B^0\|_F^2 = \mathcal{O}_{\mathbb{P}}\left(C_2^4 \frac{p^2 qs_0\log(p+q)}{n}\right), \tag{4.7}$$

where the constant C_2 comes from Lemma 2.2.

REMARK 3. If $C_2^2 = O(\eta)$, a comparison with the rates obtained in Koltchinskii, Lounici and Tsybakov (2011) shows that the rates agree. In contrast to the rate obtained for the Huber loss [equation (4.4)] the distribution of the errors is assumed to have a density. This leads to a constant of order $O(\eta^2)$ in a "worst-case" scenario. It is a natural consequence of the stronger assumption on the distribution of the errors. This is comparable to the constant obtained in Koltchinskii, Lounici and Tsybakov (2011).

In an analogy to the previous case, we are able to derive a rate for the estimation error measured in nuclear norm:

$$\|\hat{B} - B^0\|_{\text{nuclear}} = \mathcal{O}_{\mathbb{P}}\left(C_2^2 pqs_0\sqrt{\frac{\log(p+q)}{nq}}\right). \tag{4.8}$$

The rates are indeed very slow but this is not surprising given that per entry the number of observations is about $n/(pq)$. The price to pay for the estimation of the reduced number of parameters ps_0 is given by the term $\log(p+q)$.

4.2. *Weak sparsity.* In what follows, the asymptotic behavior of the proposed estimators is discussed when applied to an estimation problem where one aims at estimating a matrix that is not exactly low-rank. With Lemma 4.1 and the rates given in the previous section, we are able to derive an explicit rate also for the approximately low-rank case. For this purpose, we assume that equation (4.1) holds.

4.2.1. *Huber estimator.* The following corollary gives rates for the estimation error of the Huber estimator when used for estimation of a not exactly low-rank matrix.

COROLLARY 4.1. *With* $q\log(1+q) = o(\frac{n}{\log(p+q)})$, *we choose*

$$\lambda_\varepsilon \asymp \sqrt{\frac{\log(p+q)}{nq}}.$$

We then have

$$\|\hat{B}_H - B^0\|_F^2 = \mathcal{O}_{\mathbb{P}}\left((\eta+\kappa)^2 C_1^4 \frac{p^2 q\log(p+q)}{n}\right)^{1-r}\rho_r^r.$$

4.2.2. *Absolute value estimator.* Using the oracle inequality under the weak sparsity assumption, we obtain the following result.

COROLLARY 4.2. *With* $q\log(1+q) = o(\frac{n}{\log(p+q)})$, *we choose*

$$\lambda_\varepsilon \asymp \sqrt{\frac{\log(p+q)}{nq}}.$$

Then we have for the Frobenius norm of the estimation error

$$\|\hat{B} - B^0\|_F^2 = \mathcal{O}_{\mathbb{P}}\left(\frac{p^2 q\log(p+q)}{n}\right)^{1-r}\rho_r^r. \tag{4.9}$$

5. Simulations. In this section, the robustness of the Huber estimator 1.7 is empirically demonstrated. In Section 5.3, the Huber estimator is compared with the estimator proposed in Klopp, Lounici and Tsybakov (2016) under models 1.4 and 1.9 with each Student t and standard Gaussian noise. The sample size ranges in all simulations for all dimensions considered here from $3p\log(p)s_0$ to pq. Between minimal and maximal sample size, there are in each case 10 points. To illustrate the rate derived in Section 4, we compute the error $\|\hat{B}_H - B^0\|_F^2$ for different dimensions of the problem under increasing number of observations.

To compute the solution of the optimization problem 1.7, functions from the Matlab library `cvx` [CVX Research (2012)] were used.

Throughout this section, the error is assumed to have the following shape:

$$\text{Error} = \frac{1}{pq}\|\hat{B}_H - B^0\|_F^2 = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q}(\hat{B}_{H_{ij}} - B^0_{ij})^2. \tag{5.1}$$

To verify the robustness of the estimator 1.7 and the rate of convergence that was derived in Section 4, we use the Student t distribution with 3 degrees of freedom. Every point in the plots corresponds to an average of 25 simulations. The value of the tuning parameter is set to

$$\lambda = 2\sqrt{\frac{\log(p+q)}{nq}}.$$

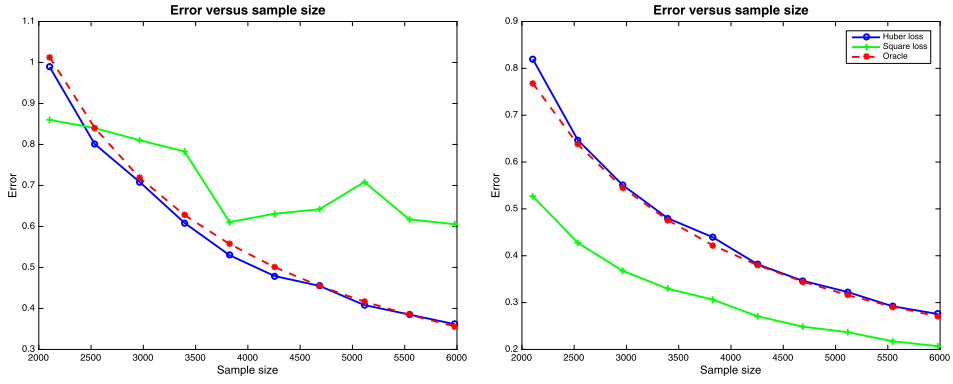
A comparison with λ_ε from Corollary 3.1 indicates that λ is rather small. For the settings we consider in this section, we found that this value for λ is more appropriate. As done in Candès and Plan (2010), for a better comparison between the error curves of our estimator and the oracle rate in equation (4.4), this rate was multiplied with 1.68 in the case of Student t distributed errors and with 1.1 in the case of Gaussian errors.

5.1. *t-distributed and Gaussian errors*. The variance of the Student t distribution with $\nu > 2$ degrees of freedom is given by

$$\text{Var}(\varepsilon_i) = \frac{\nu}{\nu - 2}, \qquad \text{for } \varepsilon_i \sim t_\nu. \tag{5.2}$$

Figure 1(a) shows a comparison between the Huber estimator 1.7 with the estimator that uses the quadratic loss in the case of Student t with 3 degrees of freedom distributed errors. As expected, the estimator that uses the quadratic loss is not robust against the corrupted entries. On the other hand, we can see in Figure 1(b) that the Huber estimator performs almost as well as the quadratic loss estimator in the case of Gaussian errors with variance 1. In agreement with the theory, the rate of the estimator is very close to the oracle rate for sufficiently large sample sizes. The value of κ that we used in the simulations is 1.345. The maximal rating η is chosen to be $\eta = 10$.

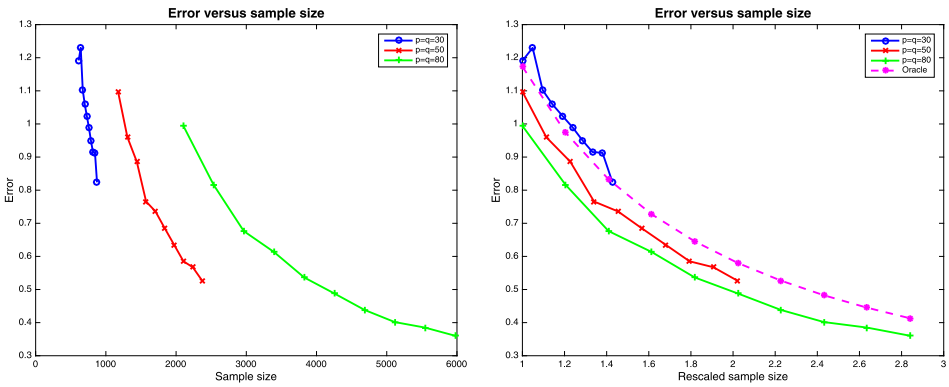
5.2. *Changing the problem size*. In order to confirm/verify the theoretical results, we proceed similar to what was done in Negahban and Wainwright (2011) and Negahban and Wainwright (2012) in the corresponding cases. Here, we consider three different problem sizes: $p, q \in \{30, 50, 80\}$. In Figure 2(a), we observe that as the problem gets harder, that is, as the dimension of the matrix increases, also the sample size needs to be larger. Figure 2(b) shows that by rescaling the sample size by $n/(3ps_0\log(p))$ the rate of convergence agrees very well with the theoretical one. It is assumed that the rank of the matrices is $s_0 = 2$ for all cases. Every point corresponds to an average of 25 simulations. The maximal rating η and the tuning parameter κ are chosen as before.



(a) (b)

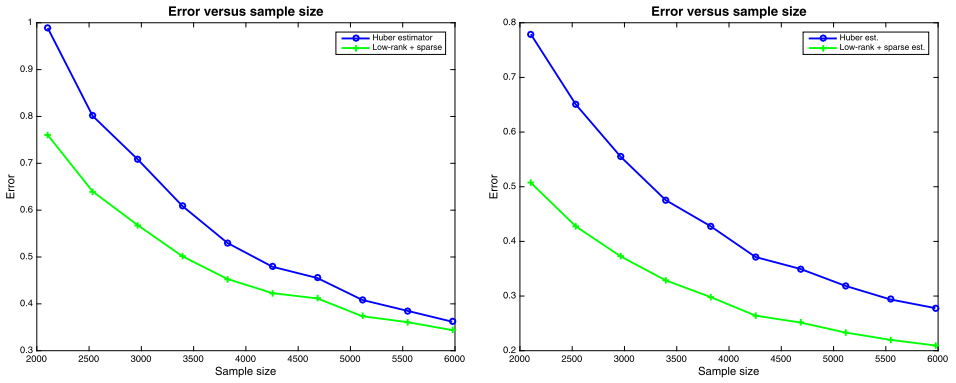
FIG. 1. *Comparison of the Huber and quadratic loss in the case of Student t distributed* (a) *and Gaussian errors* (b). *We choose* $p = q = 80$ *and* $s_0 = 2$. *The dot-dashed red line corresponds to* 1.68 *multiplied with the oracle bound derived in equation* (4.4) *for the exact low-rank case*.

5.3. *Comparison with a low-rank + sparse estimator.* In this subsection, we compare the performance of the Huber estimator 1.7 with the performance of the low-rank matrix estimator proposed by Klopp, Lounici and Tsybakov (2016) 1.10. We first compare the estimators $\hat{B}_H$ and $\hat{L}$ with the observations Y_i generated according to the model 1.4 with standard Gaussian and Student t with 3 degrees of freedom distributed errors. Equation (21) in Klopp, Lounici and Tsybakov (2016)



(a) (b)

FIG. 2. *Three different problem sizes are considered*: $p = q \in \{30, 50, 80\}$. *The rank is fixed to* $s_0 = 2$ *in both cases. In Panel* (b), *it can be seen that the rate of convergence corresponds approximately to the theoretical one derived in equation* (4.4). *The dot-dashed line is the oracle. It was multiplied by* 1.68 *in order to fit our curves*.



(a) (b)

FIG. 3. *The left panel shows the Huber estimator 1.7 and the low-rank plus sparse estimator 1.10 under the model 1.4 with Student t noise with 3 degrees of freedom. The right panel shows the same estimators under the same model with standard Gaussian noise.*

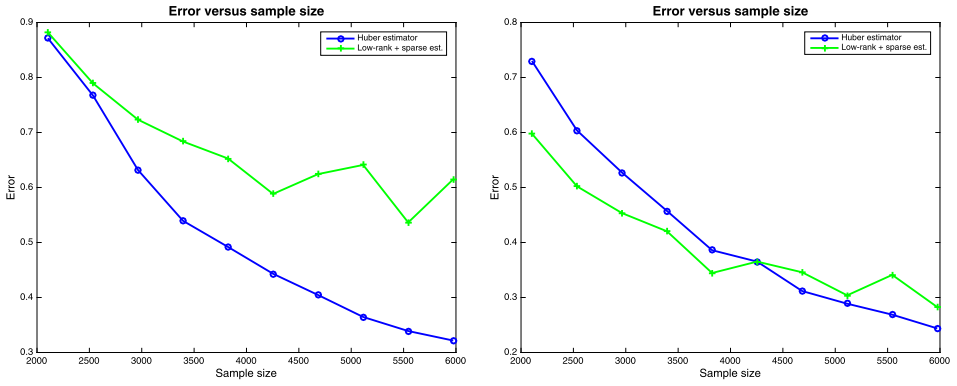
suggests that the tuning parameters are chosen as follows:

$$\lambda_1 = 2\sqrt{\frac{\log(p+q)}{nq}}, \qquad \lambda_2 = 2\frac{\log(p+q)}{n},$$

where λ_1 and λ_2 are the tuning parameters of the estimator 1.10. Also in this case it has to be noticed that the tuning parameters are smaller than the theoretical values given in their paper.

In Figure 3(a), the Huber estimator 1.7 is compared with the low-rank plus sparse estimator 1.10 under the model 1.4 with i.i.d. Student t noise with 3 degrees of freedom. As expected, these estimators perform comparably well under the trace regression model 1.4. In Figure 3(b), the same estimators are compared under the model 1.4 with i.i.d. standard Gaussian noise. Also in this case, we see that both estimators achieve approximately the same error. These observations are not surprising since the theoretical analysis of Section 3 could be carried over by adapting the (semi-)norms to the different penalization.

We now consider the model proposed in Klopp, Lounici and Tsybakov (2016) where around 5% of the observed entries are taken to be only one rating. This is the case of malicious users who systematically rate only one particular movie with the same rating. We refer to Section 2.3 of Klopp, Lounici and Tsybakov (2016) for more details on this setting. In Figure 4(a), we see that the Huber estimator outperforms the low-rank plus sparse estimator with Student t noise with 3 degrees of freedom. This might be due to the quadratic loss function and to the choice of the tuning parameters. In Figure 4(b) where Gaussian noise is considered, we observe that both estimators perform almost equally well.



(a) (b)

FIG. 4. *The left panel shows the Huber estimator* 1.7 *and the low-rank plus sparse estimator* 1.10 *under the model* 1.9 *with Student t noise with* 3 *degrees of freedom. The right panel shows the same estimators under the same model as on the left panel but with standard Gaussian noise.*

6. Discussion. In this paper, we have derived sharp and nonsharp oracle inequalities for two robust nuclear norm penalized estimators of the noisy matrix completion problem. The robust estimators were defined using the well-known Huber loss for which the sharp oracle inequality has been derived and the absolute value loss for which we have shown a nonsharp oracle inequality. For both types of oracle inequalities, we proved a general deterministic result first and added then the part arising from the empirical process. We have also shown how to apply the oracle inequalities to the case where we only assume weak sparsity, that is, approximately low-rank matrices. It is worth pointing out that our estimators do not require the distribution on the set of matrices (1.5) to be known in contrast to, for example, Koltchinskii, Lounici and Tsybakov (2011). In our case, the distribution on the set of matrices (1.5) is only needed in the theoretical analysis. The proofs of the oracle inequalities rely on the properties of the nuclear norm, and for the empirical process part on the concentration, symmetrization and contraction theorems. A main tool in this context was also the bound on the largest singular value of a matrix with finite Orlicz norm. Our simulations, in the case of the Huber loss, showed a very good agreement with the convergence rates obtained by our theoretical analysis. We saw that the oracle rate is attained up to constants in presence of non-Gaussian noise and that the robust estimation procedure outperforms the quadratic loss function.

It is left to future research to establish a sharp oracle inequality also for the case of a nondifferentiable robust loss function. The contraction inequality used in this paper for the Huber loss requires that also the derivative of the loss is Lipschitz continuous. This is not the case for the absolute value loss. Thanks to the convexity of the loss function it might be possible to derive a sharp result also for this case.

APPENDIX A: PROOFS OF MAIN RESULTS

PROOF OF INEQUALITY 2.8. Using the triangle property at B^+ with $B' = B^+$, we obtain

$$\begin{aligned} 0 &= \|B^+\|_{\text{nuclear}} - \|B^+\|_{\text{nuclear}} \\ &\le \underbrace{\Omega^+(B^+ - B^+)}_{=0} - \Omega^-(B^+) \\ &\Rightarrow \Omega^-(B^+) = 0. \end{aligned}$$

By the triangle property at B^+ with $B' = B = B^+ + B^-$, we have that

$$\begin{aligned} &\|B^+\|_{\text{nuclear}} - \|B\|_{\text{nuclear}} \\ &\quad = \|B^+\|_{\text{nuclear}} - \|B^+ + B^-\|_{\text{nuclear}} \\ &\quad \le \Omega^+(B^-) - \Omega^-(B^+ + B^-) \\ &\quad = -\Omega^-(B^+ + B^-). \end{aligned}$$

By the triangle inequality, it follows using $\Omega^-(B^+) = 0$ that

$$\Omega^-(B^+ + B^-) \ge \Omega^-(B^-) - \Omega^-(B^+) = \Omega^-(B^-).$$

Therefore, we have

$$\|B^+\|_{\text{nuclear}} - \|B^+ + B^-\|_{\text{nuclear}} \le -\Omega^-(B^-),$$

and by the triangle inequality

$$\|B^+\|_{\text{nuclear}} - \|B^+ + B^-\|_{\text{nuclear}} \ge -\|B^-\|_{\text{nuclear}},$$

which gives

$$\Omega^-(B^-) \le \|B^-\|_{\text{nuclear}}.$$

For an arbitrary B, we have again by the triangle inequality,

$$\|B\|_{\text{nuclear}} - \|B'\|_{\text{nuclear}} \le \|B^+\|_{\text{nuclear}} + \|B^-\|_{\text{nuclear}} - \|B'\|_{\text{nuclear}}.$$

Applying the triangle property at B^+, we find that

$$\|B\|_{\text{nuclear}} - \|B'\|_{\text{nuclear}} \le \Omega^+(B^+ - B') - \Omega^-(B') + \|B^-\|_{\text{nuclear}}.$$

Apply now twice the triangle inequality (first inequality) to find that

$$\begin{aligned} \|B\|_{\text{nuclear}} - \|B'\|_{\text{nuclear}} &\le \Omega^+(B - B') + \Omega^+(B^-) - \Omega^-(B - B') \\ &\quad + \Omega^-(B) + \|B^-\|_{\text{nuclear}} \\ &\le \Omega^+(B - B') - \Omega^-(B - B') + 2\|B^-\|_{\text{nuclear}}, \end{aligned}$$

where it was used that $\Omega(B^-) = 0$ and that $\Omega^-(B) \le \Omega^-(B^-) \le \|B^-\|_{\text{nuclear}}$. □

PROOF OF LEMMA 3.1. Let $B \in \mathcal{B}$. Define for $0 < t < 1$

$$\tilde{B}_t := (1-t)\hat{B} + tB.$$

Since $\mathcal{B}$ is convex, we have that $\tilde{B}_t \in \mathcal{B}$ for all $0 < t < 1$. Since $\hat{B}$ is the minimizer of the objective function and by the convexity of the objective function, we have

$$\begin{aligned} R_n(\hat{B}) + \lambda \|\hat{B}\|_{\text{nuclear}} &\leq R_n(\tilde{B}_t) + \lambda \|\tilde{B}_t\|_{\text{nuclear}} \\ &\leq R_n(\tilde{B}_t) + (1-t)\lambda \|\hat{B}\|_{\text{nuclear}} + t\lambda \|B\|_{\text{nuclear}}. \end{aligned}$$

Finally, we can conclude that

$$\frac{R_n(\hat{B}) - R_n(\tilde{B})}{t} \leq \lambda \|B\|_{\text{nuclear}} - \lambda \|\hat{B}\|_{\text{nuclear}}.$$

Letting $t \to 0$ the claim follows. □

PROOF OF THEOREM 3.1. The first-order Taylor expansion of R at $\hat{B}$ is given by

$$R(B) = R(\hat{B}) + \operatorname{trace}\big(\dot{R}(\hat{B})^T (B - \hat{B})\big) + \operatorname{Rem}(\hat{B}, B). \tag{A.1}$$

Then it follows that

$$R(\hat{B}) - R(B) + \operatorname{Rem}(\hat{B}, B) = -\operatorname{trace}\big(\dot{R}(\hat{B})^T (B - \hat{B})\big). \tag{A.2}$$

Case 1. If

$$\begin{aligned} &\operatorname{trace}\big(\dot{R}(\hat{B})^T (B - \hat{B})\big) \\ &\quad \geq \delta \underline{\lambda} \Omega^+(\hat{B} - B) + \delta \underline{\lambda} \Omega^-(\hat{B} - B) - 2\lambda \big\|B^-\big\|_{\text{nuclear}} - \lambda_*, \end{aligned} \tag{A.3}$$

then by the two-point-margin condition 3 we find that

$$R(B) - R(\hat{B}) \geq \operatorname{trace}\big(\dot{R}(\hat{B})^T (B - \hat{B})\big) + G\big(\|B - \hat{B}\|_F\big), \tag{A.4}$$

which implies that

$$\begin{aligned} R(B) - R(\hat{B}) &\geq \delta \underline{\lambda} \Omega^+(\hat{B} - B) + \delta \underline{\lambda} \Omega^-(\hat{B} - B) - 2\lambda \big\|B^-\big\|_{\text{nuclear}} \\ &\quad - \lambda_* + \underbrace{G\big(\|B - \hat{B}\|_F\big)}_{\geq 0} \\ &\geq \delta \underline{\lambda} \Omega^+(\hat{B} - B) + \delta \underline{\lambda} \Omega^-(\hat{B} - B) - 2\lambda \big\|B^-\big\|_{\text{nuclear}} - \lambda_*. \end{aligned}$$

Case 2. Assume in the following that

$$\begin{aligned} &\operatorname{trace}\big(\dot{R}(\hat{B})^T (B - \hat{B})\big) \\ &\quad \leq \delta \underline{\lambda} \Omega^+(\hat{B} - B) + \delta \underline{\lambda} \Omega^-(\hat{B} - B) - 2\lambda \big\|B^-\big\|_{\text{nuclear}} - \lambda_*. \end{aligned} \tag{A.5}$$

By the two-point inequality (Lemma 3.1), we have that

$$-\operatorname{trace}\big(\dot{R}_n(\hat{B})^T(B-\hat{B})\big) \le \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}}, \tag{A.6}$$

which implies that

$$0 \le \operatorname{trace}\big(\dot{R}_n(\hat{B})^T(B-\hat{B})\big) + \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}}. \tag{A.7}$$

Hence,

$$\begin{aligned}
&-\operatorname{trace}\big(\dot{R}(\hat{B})^T(B-\hat{B})\big) + \delta\underline{\lambda}\Omega^+(\hat{B}-B) + \delta\underline{\lambda}\Omega^-(\hat{B}-B)\\
&\quad\le \operatorname{trace}\big(\big(\dot{R}_n(\hat{B}) - \dot{R}(\hat{B})\big)^T(B-\hat{B})\big) + \delta\underline{\lambda}\Omega^+(\hat{B}-B) + \delta\underline{\lambda}\Omega^-(\hat{B}-B)\\
&\qquad + \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}}\\
&\quad\le \lambda_\varepsilon\underline{\Omega}(\hat{B}-B) + \lambda_* + \delta\underline{\lambda}\Omega^+(\hat{B}-B) + \delta\underline{\lambda}\Omega^-(\hat{B}-B)\\
&\qquad + \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}}\\
&\quad\le \lambda_\varepsilon\Omega^+(\hat{B}-B) + \lambda_\varepsilon\Omega^-(\hat{B}-B) + \lambda_* + \delta\underline{\lambda}\Omega^+(\hat{B}-B)\\
&\qquad + \delta\underline{\lambda}\Omega^-(\hat{B}-B) + \lambda\Omega^+(\hat{B}-B) - \lambda\Omega^-(\hat{B}-B) + 2\lambda\|B^-\|_{\text{nuclear}}\\
&\quad= \overline{\lambda}\Omega^+(\hat{B}-B) - (1-\delta)\underline{\lambda}\Omega^-(\hat{B}-B) + 2\lambda\|B^-\|_{\text{nuclear}} + \lambda_*.
\end{aligned}$$

Therefore, by equation (A.5),

$$\Omega^-(\hat{B}-B) \le \frac{\overline{\lambda}}{(1-\delta)\underline{\lambda}}\Omega^+(\hat{B}-B).$$

We then have by the convex conjugate inequality,

$$\begin{aligned}
\Omega^+(\hat{B}-B) &\le \|\hat{B}-B\|_F 3\sqrt{s}\\
&\le H(3\sqrt{s}) + G\big(\|\hat{B}-B\|_F\big),
\end{aligned}$$

which implies that

$$\begin{aligned}
&-\operatorname{trace}\big(\dot{R}(\hat{B})^T(B-\hat{B})\big) + \underline{\lambda}\Omega^-(\hat{B}-B) + \delta\underline{\lambda}\Omega^+(\hat{B}-B)\\
&\quad= R(\hat{B}) - R(B) + \operatorname{Rem}(\hat{B}, B) + \underline{\lambda}\Omega^-(\hat{B}-B) + \delta\underline{\lambda}\Omega^+(\hat{B}-B)\\
&\quad\le H(\overline{\lambda}3\sqrt{s}) + G\big(\|\hat{B}-B\|_F\big) + 2\lambda\|B^-\|_{\text{nuclear}} + \lambda_*\\
&\quad\le H(\overline{\lambda}3\sqrt{s}) + \operatorname{Rem}(\hat{B}, B) + 2\lambda\|B^-\|_{\text{nuclear}} + \lambda_*. \qquad \square
\end{aligned}$$

PROOF OF THEOREM 3.2. We start the proof with the following inequality using the fact that $\hat{B}$ is the minimizer of the objective function:

$$R_n(\hat{B}) + \lambda\|\hat{B}\|_{\text{nuclear}} \le R_n(B) + \lambda\|B\|_{\text{nuclear}}. \tag{A.8}$$

Then, by adding and subtracting $R(\hat{B})$ on the left-hand side and $R(B)$ on the right-hand side, we obtain

$$R(\hat{B}) - R(B) \leq -\big[(R_n(\hat{B}) - R(\hat{B}) - \big(R_n(B) - R(B)\big)\big] + \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}}.$$

Applying Assumption 3.1, the definition of $\underline{\Omega}$ and Lemma 2.8, we obtain

$$\begin{aligned} R(\hat{B}) - R(B) &\leq \lambda_\varepsilon \underline{\Omega}(\hat{B} - B) + \lambda_* + \lambda\|B\|_{\text{nuclear}} - \lambda\|\hat{B}\|_{\text{nuclear}} \\ &\leq \lambda_\varepsilon \Omega^+(\hat{B} - B) + \lambda_\varepsilon \Omega^-(\hat{B} - B) + \lambda_* \\ &\quad + \lambda\Omega^+(\hat{B} - B) - \lambda\Omega^-(\hat{B} - B) + 2\lambda\|B^-\|_{\text{nuclear}} \\ &= (\lambda_\varepsilon + \lambda)\Omega^+(\hat{B} - B) - (\lambda - \lambda_\varepsilon)\Omega^-(\hat{B} - B) + \lambda_* + 2\lambda\|B^-\|_{\text{nuclear}}. \end{aligned}$$

Since later on we apply Assumption 2, we subtract on both sides of the above inequality $R(B^0)$:

$$\begin{aligned} &R(\hat{B}) - R\big(B^0\big) + \underline{\lambda}\Omega^-(\hat{B} - B) \\ &\quad \leq R(B) - R\big(B^0\big) + \overline{\lambda}\Omega^+(\hat{B} - B) + \lambda_* + 2\lambda\|B^-\|_{\text{nuclear}}. \end{aligned} \tag{A.9}$$

It is then useful to make the following case distinction that allows us to obtain an upper bound for the estimation error.

Case 1. If $\overline{\lambda}\Omega^+(\hat{B} - B) \leq \frac{(1-\delta)}{\delta}(\lambda_* + R(B) - R(B^0) + 2\lambda\|B^-\|_{\text{nuclear}})$, then

$$\begin{aligned} \delta\overline{\lambda}\Omega^+(\hat{B} - B) &\leq (1 - \delta)\big(\lambda_* + R(B) - R\big(B^0\big) + 2\lambda\|B^-\|_{\text{nuclear}}\big) \\ &\leq \lambda_* + R(B) - R\big(B^0\big) + 2\lambda\|B^-\|_{\text{nuclear}}. \end{aligned}$$

By multiplying equation (A.9) on both sides with δ, we arrive at

$$\delta\underline{\lambda}\Omega^-(\hat{B} - B) \leq \lambda_* + R(B) - R\big(B^0\big) + 2\lambda\|B^-\|_{\text{nuclear}}.$$

Therefore,

$$\begin{aligned} &\delta\big(\overline{\lambda}\Omega^+(\hat{B} - B) + \underline{\lambda}\Omega^-(\hat{B} - B)\big) \\ &\quad \leq 2\lambda_* + 2\big(R(B) - R\big(B^0\big)\big) + 4\lambda\|B^-\|_{\text{nuclear}}. \end{aligned} \tag{A.10}$$

And since

$$\underline{\lambda} < \overline{\lambda},$$

we conclude that

$$\delta\underline{\lambda}\big(\Omega^+ + \Omega^-\big)(\hat{B} - B) \leq 2\lambda_* + 2\big(R(B) - R\big(B^0\big)\big) + 4\lambda\|B^-\|_{\text{nuclear}}.$$

Case 2. If $\overline{\lambda}\Omega^+(\hat{B} - B) \geq \frac{(1-\delta)}{\delta}(\lambda_* + R(B) - R(B^0) + 2\lambda\|B^-\|_{\text{nuclear}})$, then

$$R(\hat{B}) - R\big(B^0\big) + \underline{\lambda}\Omega^-(\hat{B} - B) \leq \overline{\lambda}\Omega^+(\hat{B} - B) + \overline{\lambda}\Omega^+(\hat{B} - B)\frac{\delta}{(1 - \delta)}.$$

This implies

$$(1-\delta)\big[R(\hat{B})-R\big(B^0\big)\big]+(1-\delta)\underline{\lambda}\Omega^-(\hat{B}-B)\le\overline{\lambda}\Omega^+(\hat{B}-B).$$

And finally we conclude that

$$\Omega^-(\hat{B}-B)\le\frac{\overline{\lambda}}{(1-\delta)\underline{\lambda}}\Omega^+(\hat{B}-B).$$

We then obtain using the definition of Ω^+ in Lemma 2.3

$$\begin{aligned}\Omega^+(\hat{B}-B)&\le\|\hat{B}-B\|_F 3\sqrt{s}\\&\le\big(\big\|\hat{B}-B^0\big\|_F+\big\|B-B^0\big\|_F\big)3\sqrt{s}\\&\le G\big(\big\|\hat{B}-B^0\big\|_F\big)+G\big(\big\|B-B^0\big\|_F\big)+2H(3\sqrt{s}).\end{aligned}$$

Invoking the convex conjugate inequality and Assumption 2, we get

$$\begin{aligned}&\delta\overline{\lambda}\Omega^+(\hat{B}-B)+\delta\underline{\lambda}\Omega^-(\hat{B}-B)\\&\quad\le 2H\big(\overline{\lambda}(1+\delta)3\sqrt{s}\big)+R(B)-R\big(B^0\big)+\big(R(B)-R\big(B^0\big)\big)\\&\qquad+\lambda_*+2\lambda\big\|B^-\big\|_{\text{nuclear}}\\&\quad\le 2H\big(\overline{\lambda}(1+\delta)3\sqrt{s}\big)+2\big(R(B)-R\big(B^0\big)\big)\\&\qquad+\lambda_*+2\lambda\big\|B^-\big\|_{\text{nuclear}}.\end{aligned}$$

Combining the two cases, we have for the estimation error

$$\begin{aligned}&\delta\underline{\lambda}\big(\Omega^++\Omega^-\big)(\hat{B}-B)\\&\quad\le 2H\big(\overline{\lambda}(1+\delta)3\sqrt{s}\big)+2\lambda_*\\&\qquad+2\big(R(B)-R\big(B^0\big)\big)+4\lambda\big\|B^-\big\|_{\text{nuclear}}\end{aligned}$$

and for the second claim we conclude that

$$\begin{aligned}R(\hat{B})-R(B)&\le\overline{\lambda}\Omega^+(\hat{B}-B)+\lambda_*+2\lambda\big\|B^-\big\|_{\text{nuclear}}\\&\le\frac{1}{\delta}\big[2H\big(\overline{\lambda}(1+\delta)3\sqrt{s}\big)+\lambda_*+2\big(R(B)-R\big(B^0\big)\big)\\&\qquad+2\lambda\big\|B^-\big\|_{\text{nuclear}}\big]+\lambda_*+2\lambda\big\|B^-\big\|_{\text{nuclear}}.\end{aligned}$$

□

PROOF OF LEMMA 2.1. The theoretical risk function arising from the Huber loss is given by

$$R(B)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X_i}\big[\mathbb{E}\big[\rho_H\big(Y_i-\operatorname{trace}(X_iB)\big)|X_i\big]\big]. \tag{A.11}$$

Suppose that X_i has its only 1 at entry (k, j). Then $XB = (B)_{jk}$. Define

$$\begin{aligned} r(x, B) &:= \mathbb{E}\big[\rho_H\big(Y_i - \operatorname{trace}(X_i B)\big) | X_i = x\big] \\ &= \mathbb{E}\big[\rho_H(Y_i - B_{jk})\big]. \end{aligned}$$

We notice that $\dot{r}(x, B) = \frac{dr(x,B)}{dB_{jk}} = \mathbb{E}[\frac{d\rho_H(Y_i - B_{jk})}{dB_{jk}}]$. The derivative with respect to B_{jk} of $\rho_H(Y_i - B_{jk})$ is given by

$$\dot{\rho}_H(Y_i - B_{jk}) := \begin{cases} -2(Y_i - B_{jk}) & \text{if } |Y_i - B_{jk}| \le \kappa, \\ -2\kappa & \text{if } Y_i - B_{jk} > \kappa, \\ 2\kappa & \text{if } Y_i - B_{jk} < -\kappa. \end{cases}$$

Then

$$\begin{aligned} \dot{r}(x, B) &= -2\int_{B_{jk}-\kappa}^{B_{jk}+\kappa} (y - B_{jk})\, dF(y) - 2\kappa \int_{B_{jk}+\kappa}^{\infty} dF(y) + 2\kappa \int_{-\infty}^{B_{jk}-\kappa} dF(y) \\ &= -2\int_{B_{jk}-\kappa}^{B_{jk}+\kappa} y\, dF(y) + 2B_{jk}\int_{B_{jk}-\kappa}^{B_{jk}+\kappa} dF(y) - 2\kappa\big[1 - F(\kappa + B_{jk})\big] \\ &\quad + 2\kappa F(B_{jk} - \kappa) \\ &= -2(B_{jk} + \kappa)F(B_{jk} + \kappa) + 2(B_{jk} - \kappa)F(B_{jk} - \kappa) \\ &\quad + 2\int_{B_{jk}-\kappa}^{B_{jk}+\kappa} F(y)\, dy + 2B_{jk}\big[F(B_{jk} + \kappa) - F(B_{jk} - \kappa)\big] - 2\kappa \\ &\quad + 2\kappa F(\kappa + B_{jk}) + 2\kappa F(B_{jk} - \kappa). \\ &= 2\int_{B_{jk}-\kappa}^{B_{jk}+\kappa} F(y)\, dy - 2\kappa. \end{aligned}$$

The second derivative of $r(x, B)$ with respect to B_{jk} is then given by

$$\ddot{r}(x, B) = 2\big[F(B_{jk} + \kappa) - F(B_{jk} - \kappa)\big].$$

Therefore, the Taylor expansion around B' is given by

$$r(x, B) = r\big(x, B'\big) + \dot{r}\big(x, B'\big)\big(B_{jk} - B'_{jk}\big) + \frac{\ddot{r}(x, \tilde{B})}{2}\big(B_{jk} - B'_{jk}\big)^2,$$

where $\tilde{B} \in \mathcal{B}$ is an intermediate point.

We can see that Assumption 3 holds with $G(u) = u^2/(2C_1^2 pq)$. □

PROOF OF LEMMA 2.2. For the (theoretical) risk function R arising from the absolute value loss, we have

$$R(B) = \mathbb{E}\big[R_n(B)\big] \tag{A.12}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[\big|Y_i - \operatorname{trace}(X_i B)\big|\big]. \tag{A.13}$$

Using the tower property of the conditional expectation, we obtain

$$R(B) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_i}\big[\mathbb{E}\big[|Y_i - \operatorname{trace}(X_i B)||X_i\big]\big]. \tag{A.14}$$

Suppose that X_i has its only 1 at entry (k, j). Then $XB = (B)_{jk}$. Define

$$\begin{aligned}
r(x, B) &:= \mathbb{E}\big[|Y_i - \operatorname{trace}(X_i B)||X_i = x\big] \\
&= \mathbb{E}\big[|Y_i - B_{jk}|\big] \\
&= \int_{y \ge B_{jk}} (y - B_{jk})\, dF(y) + \int_{y < B_{jk}} (B_{jk} - y)\, dF(y) \\
&= \int_{y \ge B_{jk}} (y - B_{jk})\, dF(y) + \int_{-\infty}^{\infty} (B_{jk} - y)\, dF(y) \\
&\quad - \int_{B_{jk}}^{\infty} (B_{jk} - y)\, dF(y) \\
&= 2\int_{B_{jk}}^{\infty} (y - B_{jk})\, dF(y) + \int_{-\infty}^{\infty} (B_{jk} - y)\, dF(y) \\
&= 2\int_{B_{jk}}^{\infty} (y - B_{jk})\, dF(y) + B_{jk} \underbrace{\int_{-\infty}^{\infty} dF(y)}_{=1} - \int_{-\infty}^{\infty} y\, dF(y) \\
&= 2\int_{B_{jk}}^{\infty} \big(1 - F(y)\big)\, dy + B_{jk} - \int_{-\infty}^{\infty} y\, dF(y).
\end{aligned}$$

The Taylor expansion of $r(x, B)$ around B^0, assuming that B^0 minimizes r, is given by

$$\begin{aligned}
r(x, B) &= r\big(x, B^0\big) + \dot{r}\big(x, B^0\big)\big(B_{jk} - B^0_{jk}\big) + \frac{\ddot{r}(x, \tilde{B})}{2}\big(B_{jk} - B^0_{jk}\big)^2 \\
&= r\big(x, B^0\big) + f(\tilde{B}_{jk})\big(B_{jk} - B^0_{jk}\big)^2,
\end{aligned}$$

where $\tilde{B} \in \mathcal{B}$ is an intermediate point;

$$r(x, B) - r\big(x, B^0\big) \ge \frac{1}{C_2^2}\big(B_{jk} - B^0_{jk}\big)^2$$

which means that the one point margin Condition 2 is satisfied with $G(u) = u^2/(2C_2^2 pq)$. □

Acknowledgements. We thank the Editor, an Associate Editor and three referees for their helpful and constructive suggestions that have led to a considerable improvement of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Robust low-rank matrix estimation" (DOI: 10.1214/17-AOS1666SUPP; .pdf). The supplemental material contains an application to real data sets, the proofs of the lemmas in Section 2 and a section on the bound of the empirical process part of the estimation problem.

REFERENCES

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CAMBIER, L. and ABSIL, P.-A. (2016). Robust low-rank matrix completion by Riemannian optimization. *SIAM J. Sci. Comput*. **38** S440–S460. MR3565571

CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** 11.

CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim*. **21** 572–596. MR2817479

CHEN, Y., JALALI, A., SANGHAVI, S. and CARAMANIS, C. (2013). Low-rank matrix recovery from errors and erasures. *IEEE Trans. Inform. Theory* **59** 4324–4337.

CHERAPANAMJERI, Y., GUPTA, K. and JAIN, P. (2016). Nearly-optimal robust matrix completion. Preprint. Available at arXiv:1606.07315.

CVX RESEARCH INC. (2012). CVX: Matlab Software for Disciplined Convex Programming, version 2.0. Available at http://cvxr.com/cvx.

ELSENER, A. and VAN DE GEER, S. (2018). Supplement to "Robust low-rank matrix estimation." DOI:10.1214/17-AOS1666SUPP.

FOYGEL, R., SHAMIR, O., SREBRO, N. and SALAKHUTDINOV, R. R. (2011). Learning with the weighted trace-norm under arbitrary sampling distributions. *Adv. Neural Inf. Process. Syst*. 2133–2141.

KLOPP, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* 282–303.

KLOPP, O., LOUNICI, K. and TSYBAKOV, A. B. (2016). Robust matrix completion. *Probab. Theory Related Fields* 1–42.

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist*. **39** 2302–2329. MR2906869

LAFOND, J. (2015). Low rank matrix completion with exponential family noise. *J. Mach. Learn. Res.*: *Workshop and Conference Proceedings. COLT* 2015 *Proceedings* **40** 1–20.

LI, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. *Constr. Approx*. **37** 73–99.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist*. **39** 1069–1097. MR2816348

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res*. **13** 1665–1697.

ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist*. **39** 887–930. MR2816342

SREBRO, N., RENNIE, J. and JAAKKOLA, T. S. (2004). Maximum-margin matrix factorization. In *Proceedings of the NIPS Conference* 1329–1336. Vancouver.

SREBRO, N. and SHRAIBMAN, A. (2005). Rank, trace-norm and max-norm. In *Learning Theory* 545–560. Springer, Berlin.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 267–288.

VAN DE GEER, S. (2001). Least squares estimation with complexity penalties. *Math. Methods Statist.* **10** 355–374. MR1867165

VAN DE GEER, S. (2016). *Estimation and Testing Under Sparsity*: *École d'Été de Probabilités de Saint-Flour XLV*-2015. Springer, Berlin.

SEMINAR FOR STATISTICS
ETH ZÜRICH
8092 ZÜRICH
SWITZERLAND
E-MAIL: elsener@stat.math.ethz.ch
geer@stat.math.ethz.ch