The Annals of Statistics
2019, Vol. 47, No. 6, 3157–3184
https://doi.org/10.1214/18-AOS1774

ADAPTIVE ESTIMATION OF THE RANK OF THE COEFFICIENT MATRIX IN HIGH-DIMENSIONAL MULTIVARIATE RESPONSE REGRESSION MODELS

BY XIN BING AND MARTEN H. WEGKAMP[1]

Cornell University

We consider the multivariate response regression problem with a regression coefficient matrix of low, unknown rank. In this setting, we analyze a new criterion for selecting the optimal reduced rank. This criterion differs notably from the one proposed in Bunea, She and Wegkamp (*Ann. Statist.* **39** (2011) 1282–1309) in that it does not require estimation of the unknown variance of the noise, nor does it depend on a delicate choice of a tuning parameter. We develop an iterative, fully data-driven procedure, that adapts to the optimal signal-to-noise ratio. This procedure finds the true rank in a few steps with overwhelming probability. At each step, our estimate increases, while at the same time it does not exceed the true rank. Our finite sample results hold for any sample size and any dimension, even when the number of responses and of covariates grow much faster than the number of observations. We perform an extensive simulation study that confirms our theoretical findings. The new method performs better and is more stable than the procedure of Bunea, She and Wegkamp (*Ann. Statist.* **39** (2011) 1282–1309) in both low- and high-dimensional settings.

1. Introduction.

1.1. *Background.* We study the multivariate response regression model

$$Y = XA + E \in \mathbb{R}^{n\times m}$$

with $X \in \mathbb{R}^{n\times p}$ of $\operatorname{rank}(X) = q$ and $A \in \mathbb{R}^{p\times m}$ of unknown $\operatorname{rank}(A) = r$. We assume that the entries E_{ij} of E are i.i.d. $N(0, \sigma^2)$ distributed with $\sigma^2 < \infty$. Section 5 discusses extensions to general, heavy tailed distributions of the errors E_{ij}.

Standard least squares estimation is tantamount to regressing each response on the predictors separately, thus ignoring the multivariate nature of the possibly correlated responses. In large dimensional settings (m and p are large relative to the sample size n), it is desirable to achieve a dimension reduction in the coefficient matrix A. One popular way of achieving this goal, is to find a common subset of $s \le p$ covariates that are relevant for prediction, using penalized least squares with

Received February 2018; revised August 2018.
[1]Supported in part by NSF Grant DMS-1712709.
MSC2010 subject classifications. 62H15, 62J07.
Key words and phrases. Multivariate response regression, reduced rank estimator, self-tuning, adaptive rank estimation, rank consistency, dimension reduction, oracle inequalities.

a ℓ_1/ℓ_2 (group lasso) type penalty on the regression coefficients; see, for instance, Bühlmann and van de Geer (2011), Bunea, She and Wegkamp (2012), Lounici et al. (2011), Obozinski, Wainwright and Jordan (2011), Yuan and Lin (2006) to recover the support of the set of s rows for which A is nonzero.

Reduced rank regression is a different approach to achieve necessary dimension reduction. The main premise is that A has low rank, so that we can write $A = A_1 A_2$ for a $p \times r$ matrix A_1 and a $r \times m$ matrix A_2. Then only few linear combinations $X^* = XA_1$ of X are needed to explain the variation of Y. Izenman (1975) coined the term *reduced-rank regression* for this class of models, but its history dates back to Anderson (1951). There are many works on this topic in the classical setting of *fixed* dimensions m and p, and sample size $n \to \infty$; see Anderson (1999), Rao (1978), Robinson (1973, 1974) and more recently, Anderson (2002). A comprehensive overview on reduced rank regression is given by Reinsel and Velu (1998). Only recently, the high-dimensional case has been discussed: Bunea, She and Wegkamp (2011, 2012), Giraud (2011, 2015), Negahban and Wainwright (2011), Rohde and Tsybakov (2011).

The main topic of this paper is the estimation of the unknown rank. Determination of the rank of the coefficient matrix is the first key step for the estimation of A. For known rank r, Anderson (1999) derives the asymptotic distribution of the reduced rank regression coefficient matrix estimator $\widehat{A}_r$ in the asymptotic setting with m, p fixed and $n \to \infty$. The estimator $\widehat{A}_k$ is the matrix corresponding to minimizing the squared Frobenius or Hilbert–Schmidt norm $\|Y - XB\|^2$ over all $p \times m$ matrices B of rank k and has a closed form, due to the Eckart–Young theorem (Eckart and Young (1936), Schmidt (1907)). It is crucial to have the true rank $k = r$ for obtaining a good fit for both $\|XA - X\widehat{A}_k\|^2$ and $\|A - \widehat{A}_k\|^2$. In general, however, the rank r is unknown *a priori*. The classical approach to estimate the rank r uses the likelihood ratio test; see Anderson (1951). An elementary calculation shows that this statistic coincides with Bartlett's test statistic as a consequence of the relation between reduced rank regression and canonical correlation analysis; see Anderson (1951), Rudelson and Vershynin (2010). Our main goal in this study is to develop a nonasymptotic method to estimate r that is easy to compute, adaptively from the data, and valid for any values of m, n and p, especially when the number of predictors p and the number of responses m are large. The resulting estimator of A can then be used to construct a possibly much smaller number of new transformed predictors, or the most important canonical variables based on the original X and Y; see Izenman ((2008), Chapter 6) for a historical account. Under weak assumptions on the signal, our estimate of r can be shown to be equal to r with overwhelming probability, to wit, $1 - \exp(-\theta_1 mn) - \exp(-\theta_2(m+q))$, for some positive, finite constants θ_1, θ_2, so that the selection error is small compared to the overall error in estimating A.

1.2. *Recent developments.* Bunea, She and Wegkamp (2011, 2012) proposed

$$\min_A \{\|Y - XA\|^2 + \mu \cdot \operatorname{rank}(A)\} \tag{1.1}$$

and recommended the choice $\mu = C(\sqrt{m} + \sqrt{q})^2\sigma^2$ with constant $C > 1$ for the tuning parameter μ. In particular, Bunea, She and Wegkamp (2011) established a convenient closed form for the rank$(\widehat{A}) = \widehat{r}$ of the matrix $\widehat{A}$ that minimizes criterion (1.1). They gave sufficient conditions on the level of the smallest nonzero singular value of XA to guarantee that $\widehat{r}$ consistently estimates r. The disadvantage of this method is that a value for σ^2, in addition to the tuning parameter C, is required for μ. Bunea, She and Wegkamp (2011) proposed to use the unbiased estimator

$$\widetilde{\sigma}^2 := \frac{\|Y - PY\|^2}{nm - qm} \tag{1.2}$$

based on the projection PY of Y onto the range space of X. However, this becomes problematic when $(n - q)m$ is not large enough, or even infeasible when $n = q$. Giraud (2011) introduces another estimation scheme that does not require estimation of σ^2. Unfortunately, a closed form for the minimizer as in Bunea, She and Wegkamp (2011) is lacking, and rank consistency in fact fails, as our simulations reveal in Appendix F.2 in the Supplementary Material (Bing and Wegkamp (2019)). Moreover, the procedures in both Bunea, She and Wegkamp (2011) and Giraud (2011) are rather sensitive to the choices of their respective tuning parameters involved. We emphasize that Giraud (2011) studies the error $\|X\widehat{A} - XA\|^2$ and not the rank of his estimator $\widehat{A}$.

1.3. *Proposed research.* This paper studies a third criterion,

$$\widehat{\sigma}_k^2 := \frac{\|Y - (PY)_k\|^2}{nm - \lambda k}. \tag{1.3}$$

Here, $(PY)_k = \sum_{j=1}^{k} d_j(PY)u_j v_j^T$ is the truncated singular value decomposition of PY based on the (decreasing) singular values $d_j(PY)$ of the projection PY and their corresponding singular vectors u_j, v_j. The range over which we minimize (1.3) is $\{0, 1, \ldots, K\}$ with $K = K_\lambda := \lfloor (nm - 1)/\lambda \rfloor \wedge q \wedge m$ to avoid a nonpositive denominator. The purpose of this paper is to show that this new criterion produces a consistent estimator of the rank. It turns out that the choice of the optimal tuning parameter λ involves a delicate trade-off. On the one hand, λ should be large enough to prevent overfitting, that is, prevent selecting a rank that is larger than the true rank r. On the other hand, if one takes λ too large, the selected rank will typically be smaller than r as the procedure will not be able to distinguish the unknown singular values $d_j(XA)$ from the noise for $j > s$, for some $s = s(\lambda) < r$. To effectively deal with this situation, we refine our initial procedure using our new criterion (1.3) by an iterative procedure in Section 4 that provenly finds the optimal value of λ and consequently of the estimate of r. This method does not require any data-splitting and our simulations show that it is very stable, even for general, heavy tailed error distributions. To our knowledge, it is a rare feat to have

a feasible algorithm that finds the optimal tuning parameter in a fully data-driven way, without data-splitting, and with mathematically proven guarantees.

While our main interest is to provide consistent estimators of the rank, we briefly address estimation of the mean XA, which is the principal problem in Bunea, She and Wegkamp (2011, 2012), Giraud (2011). Our selected rank $\widehat{k}$ automatically yields the estimate $X\widehat{A} := (PY)_{\widehat{k}}$. We prove in Theorem 10 in Appendix D of the Supplementary Material (Bing and Wegkamp (2019)) that on the event $\widehat{k} = r$, the inequality $\|X\widehat{A} - XA\|^2 \le 4rd_1^2(PE)$ holds, for our selected rank $\widehat{k}$. This provides a direct link between rank consistency and optimal estimation of the mean, because $d_1^2(PE) \le 2(m+q)\sigma^2$ with overwhelming probability; see (3.1) and (3.2) below. (In fact, this bound continues to hold, up to a multiplicative constant, for sub-Gaussian errors, using Theorem 5.39 of Vershynin (2012).) Hence we can estimate XA at the rate $r(m+q)$, which is proportional to the number of parameters in the low rank model and minimax optimal (Bunea, She and Wegkamp (2011, 2012), Giraud (2011)). Simulations in Appendix F.3 of the Supplementary Material (Bing and Wegkamp (2019)) show that our procedures in fact provide better estimates of XA than their competitors, even in approximately low-rank models.

The paper is organized as follows. Section 2 shows that the minimizer of (1.3) has a closed form. The main results are discussed in Sections 3 and 4. It obtains rank consistency in case of no signal ($XA = 0$) and in case of sufficient signal. For the latter, we develop a key notion of signal-to-noise ratio that is required for rank consistency. A sufficient, easily interpretable condition will be presented that corresponds to a computable value (estimate) of the tuning parameter λ. We develop in Section 4 an iterative, fully automated procedure, which has a guaranteed recovery of the true rank (with overwhelming probability) under *increasingly milder* conditions on the signal. The first step uses the potentially suboptimal estimate $\widehat{k}_0$ developed in Section 3, but which is less than r, with overwhelming probability. This value $\widehat{k}_0$ is used to update the tuning parameter λ. Then we minimize (1.3) again, and obtain a new estimate $\widehat{k}_1$, which in turn is used to update λ. The procedure produces each time a smaller λ, thereby selecting a larger rank k than the previous one, while each time we can guarantee that the selected rank does not exceed the true rank r. This is a major mathematical challenge and its proof relies on highly nontrivial monotonicity arguments. The procedure stops when the selected rank does not change after an iteration. Our results hold with high probability (exponential in mn and $m+q$) which translates into extremely accurate estimates under a weak signal condition.

Section 5 describes several extensions of the developed theory, allowing for non-Gaussian errors E_{ij}.

A large simulation study is reported in Section 6. It confirms our theoretical findings, and shows that our method improves upon the methods proposed in Bunea, She and Wegkamp (2011).

The proofs are deferred to the Supplementary Material (Bing and Wegkamp (2019)).

1.4. *Notation.* For any matrix A, we will use $A_{k\ell}$ to denote the k,ℓth element of A (i.e., the entry on the kth row and ℓth column of A), and we write $d_1(A) \geq d_2(A) \geq \cdots$ to denote its ordered singular values.

The Frobenius or Hilbert–Schmidt inner product $\langle \cdot, \cdot \rangle$ on the space of matrices is defined as $\langle A, B \rangle = \operatorname{tr}(A^T B)$ for commensurate matrices A, B. The corresponding norm is denoted by $\| \cdot \|^2$, and we recall that $\|A\|^2 = \operatorname{tr}(A^T A) = \sum_j d_j^2(A)$ for any matrix A. Moreover, it is known (Eckart and Young (1936), Schmidt (1907), see also the review by Stewart (1993)) that minimizing $\|A - B\|^2$ over B with $r(B) \leq r$ is achieved for $B = (A)_r = U D_r V^T$ based on the singular value decomposition of $A = U D V^T$ where D_r denotes the diagonal matrix with $[D]_{ii} = d_i(A)$ for $i = 1, \ldots, r$. Hence, $\min_{B: r(B) = r} \|A - B\|^2 = \sum_{j > r} d_j^2(A)$.

For other norms on matrices, we use $\| \cdot \|_2$ to denote the operator norm and $\| \cdot \|_*$ the nuclear norm (i.e., the sum of singular values). We have the inequalities $\langle A, B \rangle = \operatorname{tr}(A^T B) \leq \|A\|_2 \|B\|_*$ and $\|A\|_* \leq \sqrt{\operatorname{rank}(A)} \|A\|$.

For two positive sequences a_n and b_n, we denote by $a_n = O(b_n)$ if there exists constant $C > 0$ such that $\lim_{n \to \infty} a_n / b_n \leq C$. If $\lim_{n \to \infty} a_n / b_n \to 0$, we write $a_n = o(b_n)$.

For general $m \times n$ matrices A and B, Weyl's inequality (Weyl (1912)) implies that $d_{i+j-1}(A + B) \leq d_i(A) + d_j(B)$ for $1 \leq i, j, \leq q$ and $i + j \leq q + 1$ with $q = \min\{m, n\}$.

We denote the projection matrix onto the column space of X by P and we write $q := \operatorname{rank}(X)$ and $N := \operatorname{rank}(PY) = q \wedge m$. We set $\widehat{\sigma}_0^2 := \|Y\|^2/(nm)$, by defining $(PY)_0 := 0$ and define $\widehat{\sigma}^2 := \|E\|^2/(nm)$. Throughout the paper, we use A to denote the true coefficient matrix and r to denote its true rank.

2. Properties of the minimizer of the new criterion. At first glance, it seems difficult to describe the minimizer $\widehat{k}$ of $k \mapsto \widehat{\sigma}_k^2$ because *both* the numerator and denominator in $\widehat{\sigma}_k^2$ are decreasing in k. However, it turns out that there is a unique minimizer with a neat explicit formula. First, we characterize the comparison between $\widehat{\sigma}_i$ and $\widehat{\sigma}_j$ for $i \neq j$.

PROPOSITION 1. *Let $i, j \in \{0, 1, \ldots, K\}$ with $i < j$. Then*

$$\widehat{\sigma}_j^2 \leq \widehat{\sigma}_i^2 \quad \Longleftrightarrow \quad \frac{1}{j - i} \sum_{k=i+1}^{j} d_k^2(PY) \geq \lambda \widehat{\sigma}_j^2. \tag{2.1}$$

In particular,

$$\widehat{\sigma}_j^2 \leq \widehat{\sigma}_{j-1}^2 \quad \Longleftrightarrow \quad d_j^2(PY) \geq \lambda \widehat{\sigma}_j^2 \tag{2.2}$$

and

$$d_j^2(PY) \leq \lambda \widehat{\sigma}_j^2 \quad \Longleftrightarrow \quad d_j^2(PY) \leq \lambda \widehat{\sigma}_{j-1}^2. \tag{2.3}$$

This result and the monotonicity of the singular values $d_1(PY) \geq d_2(PY) \geq \cdots$ readily yield the following statement.

PROPOSITION 2. *Let* $k \in \{1, \ldots, K\}$. *Then*

$$\widehat{\sigma}_k^2 \leq \widehat{\sigma}_{k-1}^2 \iff \widehat{\sigma}_k^2 \leq \widehat{\sigma}_\ell^2 \qquad \text{for all } \ell \leq k-1, \tag{2.4}$$

$$\widehat{\sigma}_k^2 \geq \widehat{\sigma}_{k-1}^2 \iff \widehat{\sigma}_\ell^2 \geq \widehat{\sigma}_{k-1}^2 \qquad \text{for all } \ell > k-1. \tag{2.5}$$

It is clear that if $\widehat{\sigma}_1^2 \geq \widehat{\sigma}_0^2$, then $k = 0$ minimizes $\widehat{\sigma}_k^2$. Likewise, if $\widehat{\sigma}_K^2 \leq \widehat{\sigma}_{K-1}^2$, then $k = K$ minimizes the criterion $\widehat{\sigma}_k^2$. After a little reflexion, we see that $\widehat{\sigma}_k^2$ is minimized at the last k for which $d_k^2(PY) \geq \lambda \widehat{\sigma}_k^2$ holds. That is,

$$\widehat{k} = \max\bigl\{0 \leq k \leq K : d_j^2(PY) \geq \lambda \widehat{\sigma}_j^2 \text{ for all } j \leq k \text{ and } d_{k+1}^2(PY) < \lambda \widehat{\sigma}_{k+1}^2\bigr\} \tag{2.6}$$

minimizes $\widehat{\sigma}_k^2$ with the convention that the maximum of the empty set is 0. Properties (2.4) and (2.5) ensure that $d_j^2(PY) \geq \lambda \widehat{\sigma}_j^2$ must hold automatically for all $j \leq k$ as well as $d_\ell^2(PY) < \lambda \widehat{\sigma}_\ell^2$ for all $\ell > k$. That is, $\widehat{k}$ has an even more convenient closed form

$$\widehat{k} = \max\bigl\{0 \leq k \leq K : d_k^2(PY) \geq \lambda \widehat{\sigma}_k^2\bigr\} = \sum_{k=1}^{K} 1\bigl\{d_k^2(PY) \geq \lambda \widehat{\sigma}_k^2\bigr\}. \tag{2.7}$$

Summarizing, we have shown the following result.

THEOREM 3. *There exists a unique minimizer* $\widehat{k}$ *of* (1.3), *given by* (2.7), *such that* $\widehat{\sigma}_k^2$ *is monotone decreasing for* $k \leq \widehat{k}$, *and monotone increasing for* $k \geq \widehat{k}$.

It is interesting to compare the choice $\widehat{k}$ in (2.7) with $\widehat{r}$ in Bunea, She and Wegkamp (2011). In that paper, it is shown that (1.1) is equivalent with

$$\min_k \bigl\{\|Y - (PY)_k\|^2 + \mu k\bigr\} \tag{2.8}$$

based on the truncated singular value decomposition UD_kV^T of the projection $PY = UDV^T$ with $D_k = \operatorname{diag}(D_{11}, \ldots, D_{kk}, 0, \ldots, 0)$. Furthermore, Bunea, She and Wegkamp (2011) uses this formulation to derive a closed form for $\widehat{r}$, to wit,

$$\widehat{r} = \sum_{k \geq 1} 1\bigl\{d_k^2(PY) \geq \mu\bigr\} \tag{2.9}$$

based on the singular values $d_1(PY) \geq d_2(PY) \geq \cdots$ of the projection PY. The main difference between (2.7) and (2.9) is that $\widehat{k}$ counts the number of singular values of PY above a *variable* threshold, while $\widehat{r}$ counts the number of singular values of PY above a *fixed* threshold. Another difference is that the fixed threshold is proportional to the unknown variance σ^2, while the variable threshold is proportional to $\widehat{\sigma}_k^2$, which can be thought of as an estimate of σ^2 only for k close to r.

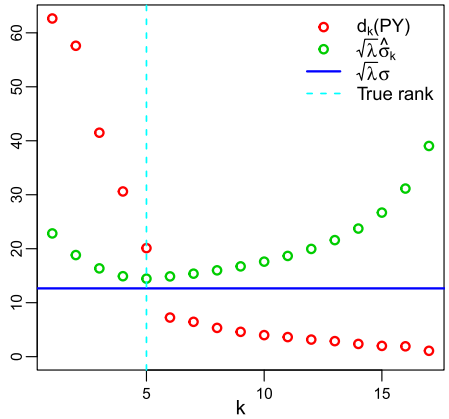

FIG. 1. *Plot of $d_k(PY)$ and $\sqrt{\lambda}\widehat{\sigma}_k$ versus k. In this experiment, we used $n = 150$, $m = 30$, $p = 20$, $q = 20$, $r = 5$, $\eta = 0.1$, $b_0 = 0.2$ and $\sigma^2 = 1$, using the notation and simulation setup of Section 6.3.*

To further illustrate the existence and uniqueness of $\widehat{k}$, we perform one experiment to show how the $d_k(PY)$ and $\widehat{\sigma}_k$ vary across different k; see Figure 1. The plot first displays the monotone property of $\widehat{\sigma}_k$ for $k \le \widehat{k}$ and $k \ge \widehat{k}$. It also justifies the definition of $\widehat{k}$ in (2.7) since the rank which minimizes $\widehat{\sigma}_k$ is exactly what we defined.

3. Rank consistency.

3.1. *The null case $XA = 0$.* We treat the case $XA = 0$ separately as the case $XA \neq 0$ requires a lower bound on the nonzero singular values of XA.

THEOREM 4. *Assume $XA = 0$. Then, on the event $\{d_1^2(PE) \le \lambda\widehat{\sigma}^2\}$, we have $\widehat{k} = 0$.*

We particularize Theorem 4 to the case where the entries of E are independent $N(0, \sigma^2)$. In that case, general random matrix theory and Borel's inequality for suprema of Gaussian processes, respectively, give

$$\mathbb{E}\big[d_1(PE)\big] \le \sigma(\sqrt{m} + \sqrt{q}) \tag{3.1}$$

and

$$\mathbb{P}\big\{d_1(PE) \ge \mathbb{E}\big[d_1(PE)\big] + \sigma t\big\} \le \exp(-t^2/2) \qquad \text{for all } t > 0; \tag{3.2}$$

see Lemma 3 in Bunea, She and Wegkamp (2011). Moreover, $(nm)\widehat{\sigma}^2$ has a central χ^2_{mn} distribution, so that general tail bounds (Johnstone (2001)) yield

$$\mathbb{P}\big\{\widehat{\sigma}^2 \le \sigma^2(1-\varepsilon)\big\} \le \exp(-mn\varepsilon^2/4), \qquad 0 \le \varepsilon < 1, \tag{3.3}$$

$$\mathbb{P}\big\{\widehat{\sigma}^2 \ge \sigma^2(1+\varepsilon)\big\} \le \exp(-3mn\varepsilon^2/16), \qquad 0 \le \varepsilon < 1/2. \tag{3.4}$$

We immediately obtain the following corollary by using (3.2)–(3.4).

COROLLARY 5. *For any* $\lambda > (\sqrt{m} + \sqrt{q})^2$, *we have* $\mathbb{P}\{\widehat{k} = 0\} \to 1$ *exponentially fast as* $nm \to \infty$ *and* $m + q \to \infty$.

3.2. *The general case* $XA \neq 0$. The range over which we minimize (1.3) is $\{0, 1, \ldots, K\}$ depends on λ as the largest possible value is

$$K = K_\lambda := \left\lfloor \frac{nm - 1}{\lambda} \right\rfloor \wedge m \wedge q \tag{3.5}$$

to avoid a nonpositive denominator in criterion (1.3).

THEOREM 6. *Assume* $r \leq K_\lambda$. *On the event*

$$d_1^2(PE) \leq \lambda \widehat{\sigma}_r^2 := \frac{\|Y - (PY)_r\|^2}{(nm/\lambda) - r}, \tag{3.6}$$

we have $\widehat{k} \leq r$. *If* $r > K_\lambda$, *then trivially* $\widehat{k} \leq r$ *holds*, *with probability one*.

The restriction $r \leq K_\lambda$ guarantees that $\widehat{\lambda}\sigma_r^2$ is positive, that is, the event (3.6) is nonempty. If $r > K_\lambda$, then $\widehat{k} \leq r$, holds trivially, with probability one, as $\widehat{k}$ is selected from $\{0, \ldots, K_\lambda\}$.

While the quantity $\widehat{\sigma}_r^2$ is a natural one in this problem, it depends on the unknown rank r. It turns out that quantifying $\widehat{\sigma}_r^2$ is not trivial.

PROPOSITION 7. *Assume* $r \leq K_\lambda$. *On the event*

$$2d_1^2(PE) \leq \lambda \widehat{\sigma}^2, \tag{3.7}$$

we have

$$\widehat{\sigma}^2 \leq \widehat{\sigma}_r^2 \leq \frac{nm}{nm - \lambda r} \widehat{\sigma}^2. \tag{3.8}$$

This result combined with Theorem 6 tells us that if λ is chosen large enough, we can guarantee that $\widehat{k} \leq r$. Moreover, this choice is independent of both r and σ^2. Indeed, for matrices E with independent $N(0, \sigma^2)$ Gaussian entries, any choice $\lambda > 2(\sqrt{m} + \sqrt{q})^2$ suffices.

THEOREM 8. *For* $\lambda = C(\sqrt{m} + \sqrt{q})^2$ *with any numerical constant* $C > 2$, $\mathbb{P}\{\widehat{k} \leq r\} \to 1$ *as* $mn \to \infty$ *and* $m + q \to \infty$.

The convergence rate in Theorem 8 is exponentially fast in nm and $m + q$. Again, if $r > K_\lambda$, then $\mathbb{P}\{\widehat{k} \leq r\} = 1$ holds trivially.

Consistency of $\widehat{k}$ can be achieved under a suitable signal to noise condition.

THEOREM 9. *For* $1 \leq s \leq r \leq K_\lambda$, *on the event*

$$d_s(XA) \geq d_1(PE) + \sqrt{\lambda}\widehat{\sigma}_r \tag{3.9}$$

intersected with the event (3.6), *we further have* $\widehat{k} \in [s, r]$.

This theorem, combined with Proposition 7, immediately yields the following corollary.

COROLLARY 10. *For $1 \le s \le r \le K_\lambda$, on the event*

$$\{2d_1^2(PE) \le \lambda\widehat{\sigma}^2\} \cap \left\{d_s(XA) \ge \sqrt{\lambda}\widehat{\sigma}\left[\frac{\sqrt{2}}{2} + \sqrt{\frac{nm}{nm-\lambda r}}\right]\right\},$$

we have $\widehat{k} \in [s, r]$.

The choice of λ impacts the possible values for $\widehat{k}$, the minimizer of criterion (1.3), and we see that the range $\{0, 1, \ldots, K_\lambda\}$ increases as λ decreases. If the true rank is rather large ($r > K_\lambda$), then no guarantees for $\widehat{k}$ can be made, except for the trivial, yet important observation that $\widehat{k} \le r$. On the other hand, if $r < K_\lambda$, which is arguably the more interesting case for *low rank* regression, then consistency guarantees can be made under a suitable condition on the rth singular value $d_r(XA)$ of XA. This condition becomes milder if λ decreases.

Let $\delta > 0$. A slightly stronger restriction for the upper bound on r,

$$r < \frac{\delta}{1+\delta}\frac{nm}{\lambda} \wedge m \wedge q \tag{3.10}$$

translates into a bound for the ratio

$$\frac{nm}{nm-\lambda r} \le 1 + \delta \tag{3.11}$$

appearing in the lower bound for the *signal* $d_s(XA)$. We can further particularize to the Gaussian setting.

THEOREM 11. *Let $\lambda = 2C(\sqrt{m} + \sqrt{q})^2$ for some numerical constant $C > 1$. Assume further that r and δ satisfy (3.10) and*

$$d_s(XA) \ge C'\sigma(\sqrt{m} + \sqrt{q}) \tag{3.12}$$

for some $s \le r$ and some numerical constant $C' > \sqrt{C}(1 + \sqrt{2(1+\delta)})$. Then $\mathbb{P}\{s \le \widehat{k} \le r\} \to 1$ as $mn \to \infty$ and $m + q \to \infty$.

In particular, if (3.12) holds for $s = r$, then $\widehat{k}$ consistently estimates r.

The convergence rate in Theorem 11 is exponentially fast in nm and $m + q$. This shows that the above procedure is highly accurate, which is confirmed in our simulation study. From the oracle inequality in the Supplementary Material (Bing and Wegkamp (2019)), the fit $\|X\widehat{A}_{\widehat{k}} - XA\|^2$, for $s \le \widehat{k} \le r$, differs only within some constant levels of the *noise level* $d_1^2(PE)$.

4. Self-tuning procedure. From Theorem 6 in Section 3, we need to take λ, or in fact $\lambda\widehat{\sigma}_r^2$, as $\widehat{\sigma}_r^2$ depends on λ, large enough to prevent overfitting, that is, to avoid selecting a $\widehat{k}$ larger than the true rank r. On the other hand, we would like to keep λ small to be able to detect a small signal level. Indeed, (3.6) in Theorem 6 states that we need

$$d_1(PE) \le \sqrt{\lambda}\widehat{\sigma}_r = \sqrt{\frac{\|Y-(PY)_r\|^2}{nm/\lambda - r}}. \tag{4.1}$$

The term on the right is decreasing in λ, so we should choose λ as small as possible such that $\sqrt{\lambda}\widehat{\sigma}_r$ is close to $d_1(PE)$. Without any prior knowledge on r, it is difficult to find the optimal choice for λ. However, Theorem 6 and Proposition 7 tell us that an initial λ_0 satisfying $\{2d_1^2(PE) \le \lambda_0\widehat{\sigma}^2\}$ yields an estimated rank $\widehat{k}_0$ with $\widehat{k}_0 \le r$. Our idea is to use this lower bound $\widehat{k}_0$ for r to reduce our value λ_0 to λ_1. This, in turn, will yield a possibly larger estimated rank $\widehat{k}_1$, which still obeys $\widehat{k}_1 \le r$. More precisely, we propose the following *Self-Tuning Rank Selection* (STRS) procedure. Let Z be a $q \times m$ matrix with i.i.d. standard Gaussian entries and define $S_j = \mathbb{E}[d_j^2(Z)]$ with the convention $S_j := 0$ for $j > N = q \wedge m$. Moreover, let $\widehat{K}_t := (nm/\widehat{\lambda}_t) \wedge N$ for given $\widehat{\lambda}_t$. For any $\varepsilon \in (0,1)$, we define

$$\widehat{\lambda}_0 := 2(1+\varepsilon)S_1, \tag{4.2}$$

$$\widehat{k}_0 := \operatorname*{arg\,min}_{0\le k\le \widehat{K}_0} \frac{\|Y-(PY)_k\|^2}{nm - \widehat{\lambda}_0 k} \tag{4.3}$$

as starting values, and if $\widehat{k}_0 \ge 1$, for $t \ge 0$, we update

$$\widehat{\lambda}_{t+1} := \frac{nm}{(1-\varepsilon)\widehat{R}_t/\widehat{U}_t + \widehat{k}_t}, \tag{4.4}$$

$$\widehat{k}_{t+1} := \operatorname*{arg\,min}_{\widehat{k}_t\le k\le \widehat{K}_{t+1}} \frac{\|Y-(PY)_k\|^2}{nm - \widehat{\lambda}_{t+1} k} \tag{4.5}$$

where

$$\widehat{R}_t := (n-q)m + \sum_{j=2\widehat{k}_t+1}^{N} S_j, \qquad \widehat{U}_t := S_1 \vee (S_{2\widehat{k}_t+1} + S_{2\widehat{k}_t+2}). \tag{4.6}$$

The procedure stops when $\widehat{k}_{t+1} = \widehat{k}_t$. The entire procedure is free of σ^2 and both $\widehat{R}_t$ and $\widehat{U}_t$ can be numerically evaluated by Monte Carlo simulations. Alternatively, we provide an analogous procedure with analytical expressions in the Supplementary Material (Bing and Wegkamp (2019)), but its performance in our simulations is actually slightly inferior to the original procedure that utilizes Monte Carlo simulations.

Regarding the computational complexity, we emphasize that the above STRS procedure has almost *the same* level of computational complexity as the methods in

(1.1) and (1.3). This is due to the fact that the computationally expensive singular value decomposition only needs to be computed once. Additionally, in order to find the new rank in step (4.5), we only need to consider values that are larger than (or equal to) the previously selected rank. This avoids a lot of extra computation.

The following proposition is critical for the feasibility of STRS.

PROPOSITION 12. *We have* $\widehat{\lambda}_t > \widehat{\lambda}_{t+1}$ *and* $\widehat{k}_t \le \widehat{k}_{t+1}$, *for all* $t \ge 0$. *More importantly,* $\widehat{k}_t \le r$ *for all* $t \ge 0$, *holds with probability tending to* 1 *exponentially fast as* $(q \vee m) \to \infty$ *and* $nm \to \infty$.

The increasing property of $\widehat{k}_t$ immediately yields the following theorem.

THEOREM 13. *Let numerical constant* $C > 2$. *Let* $\widetilde{k}$ *be the minimizer of* (1.3) *using* $\lambda = C(\sqrt{m} + \sqrt{q})^2$ *and* $\widehat{k}$ *be the final selected rank of STRS starting from the same value* $\widehat{\lambda}_0 = \lambda$. *Then*

$$\mathbb{P}\{\widetilde{k} \le \widehat{k} \le r\} \to 1,$$

as $nm \to \infty$ *and* $(q \vee m) \to \infty$.

We find that STRS always selects a rank closer to the true rank than the (one step) Generalized Rank Selection Procedure (GRS) from the previous section that uses (1.3) as its criterion.

The decreasing property of $\widehat{\lambda}_t$, stated in Proposition 12, implies an increasingly milder condition on the required signal. Meanwhile, the way of updating λ in step (4.4) is carefully chosen to maintain $\widehat{k}_t \le r$. We refer to the proof for more explanations. Thus, if a proper sequence of signal-to-noise condition is met, we expect that STRS finds the rank consistently. The following theorem confirms this.

THEOREM 14. *Let* $k_0 < k_1 < \cdots < k_T = r$ *be a strictly increasing subsequence of* $\{1, 2, \ldots, r\}$ *of length* $T + 1 \le r$. *Define* λ_0 *as* (4.2) *and* λ_{t+1} *obtained from* (4.4) *by using* k_t *in lieu of* $\widehat{k}_t$, *for* $t = 0, \ldots, T - 1$. *Assume* r *and* δ *satisfy* (3.10) *for* λ_0. *Then, on the event*

$$d_{k_t}(XA) \ge C''\sigma\sqrt{\lambda_t}, \qquad t = 0, \ldots, T \tag{4.7}$$

for some numerical constant $C'' > 1/\sqrt{2} + \sqrt{1+\delta}$, *there exists* $0 \le T' \le T$ *such that* $\widehat{k}_{T'} = r$, *with probability tending to* 1, *where* $\widehat{k}_0 \le \widehat{k}_1 \le \cdots \le \widehat{k}_{T'}$ *are from* (4.5).

The sequence of $\{k_0, \ldots, k_T\}$ plays an important role for interpreting the above theorem. It can be regarded as a underlying sequence bridge starting from 1 and leading toward the true rank. The ideal case is $\{k_0, \ldots, k_T\} = \{r\}$ which leads to a one-step recovery, but requires a comparatively stronger signal-to-noise condition (4.7). At the other extreme, it could take r steps to recover the true rank. We emphasize that the latter case requires the *mildest* signal-to-noise condition by the following two observations:

(1) *Proposition 12 guarantees that each time the updated* λ_t *is decreasing*;
(2) *The signal condition* (4.7) *is becoming milder as* λ_t *gets smaller*.

From display (3.8) in Proposition 7 and from Corollary 10, it is clear that we use the string of inequalities $\widehat{\sigma}^2 \le \widehat{\sigma}_r^2 \le nm/(nm - \lambda r)\widehat{\sigma}^2$ to derive the required signal-to-noise condition. The second inequality becomes loose for large r such that $nm - \lambda r$ is small, or equivalently, δ is large from (3.11), which further implies a possible larger decrement of the required signal-to-noise condition by using STRS. To illustrate this phenomenon concretely, we study to a special case where $r \ge N/2$ and $\{k_0, \dots, k_T\} = \{\lceil N/2 \rceil, r\}$, and show in the theorem below that the signal condition for recovering r can be relaxed significantly when δ is large.

THEOREM 15. *Define* $\lambda_0 = 2C(\sqrt{m} + \sqrt{q})^2$ *with* $C = 8/7$. *Assume* r *and* δ *satisfy* (3.10) *for* λ_0 *and*

$$d_{\lceil N/2 \rceil}(XA) \ge C'\big[1 + \sqrt{2(1+\delta)}\big]\sigma(\sqrt{m} + \sqrt{q}), \tag{4.8}$$

$$d_r(XA) \ge C'\left[1 + \sqrt{\frac{1+\delta}{1+\delta/8}}\right]\sigma(\sqrt{m} + \sqrt{q}) \tag{4.9}$$

for some numerical constant $C' > 2\sqrt{2/7}$. *Then either* $\widehat{k}_0 = r$ *or* $\widehat{k}_1 = r$, *with probability tending to* 1, *as* $N = q \wedge m \to \infty$. *Here,* $\widehat{k}_0$ *and* $\widehat{k}_1$ *are selected from* (4.3) *and* (4.5).

The lower bound condition (4.8) is condition (3.12) with $s = \lceil N/2 \rceil < r$. As a simple numerical illustration, we compare (4.8) (and, therefore, (3.12)) with (4.9) for $\delta = 4$ and 100. If $\delta = 4$, we obtain

$$d_{\lceil N/2 \rceil}(XA) \ge 4.17C'(\sqrt{m} + \sqrt{q})\sigma, \qquad d_r(XA) \ge 2.83C'(\sqrt{m} + \sqrt{q})\sigma,$$

while if $\delta = 100$, we have

$$d_{\lceil N/2 \rceil}(XA) \ge 15.3C'(\sqrt{m} + \sqrt{q})\sigma, \qquad d_r(XA) \ge 3.74C'(\sqrt{m} + \sqrt{q})\sigma.$$

As we can see, for small δ, the signal-to-noise condition decreases slightly. However, for larger δ, $\sqrt{1+\delta}$ could be quite large, while $\sqrt{(1+\delta)/(1+\delta/8)}$ is always bounded above by $2\sqrt{2}$. To further elaborate the implications of small/large δ, we consider two cases by recalling the rank constraint (3.10):

(1) If $nm/\lambda_0 \ge (1+\delta)N/\delta$, the rank constraint (3.10) reduces to simply $r \le N$. From (3.11), a smaller value for δ leads to a smaller value for $nm/(nm - \lambda_0 r)$, provided $r \le N$. Therefore, $\widehat{\sigma}^2 \le \widehat{\sigma}_r^2 \le nm/(nm - \lambda_0 r)\widehat{\sigma}^2$ should be tight and we expect a smaller reduction of the signal condition for smaller values of δ. On the other hand, when nm and $\lambda_0 N$ are close, meaning δ is large, we expect a considerable relaxation of the signal condition for comparatively large r. These two points are clearly reflected in (4.8) and (4.9).

(2) If $nm/\lambda_0 \le (1+\delta)N/\delta$, it follows that $r \le \{\delta/(1+\delta)\}(nm/\lambda_0)$. Then a smaller δ means a stronger restriction on r which implies $\widehat{\sigma}^2 \le \widehat{\sigma}_r^2 \le nm/(nm - \lambda_0 r)\widehat{\sigma}^2$ becomes tight and we expect a modest relaxation of the signal condition. If δ is large, then $nm - \lambda_0 r$ could be small for a comparatively large r. Thus $nm/(nm - \lambda_0 r)$ would explode and we expect a significant decrease in the lower bound (4.9) for the signal. Both these observations agree with our results. It is worth mentioning that when nm is small comparing to $\lambda_0 N$, δ is likely to be large. For instance, when $m = q = n$ is moderate, taking $\lambda_0 = 2(\sqrt{m} + \sqrt{q})^2$ yields $nm/\lambda_0 = q/8$ which is not quite large already. Imposing a small δ in this case would further restrict the range of r.

Recall that the range of allowable rank $\{0, \ldots, K_\lambda\}$ in (3.5) increases as λ decreases. This means that, after a few iterations, the true rank could be selected even when it was out of the possible range $\{0, \ldots, K_{\lambda_0}\}$ at the beginning. This phenomenon is clearly supported by our simulations in Section 6.5. In addition, the following proposition proves that K_{λ_0} can be extended to $N = q \wedge m$ in some settings even when (3.10) is not met for λ_0.

PROPOSITION 16. *Let* $\lambda_0 = 2C(\sqrt{m} + \sqrt{q})^2$ *with* $C = 8/7$ *and assume* $nm/\lambda_0 \ge 3N/4$. *Suppose the first selected rank from* (4.3) *by using* λ_0 *satisfies* $\widehat{k}_0 \ge N/2$ *and*

$$d_r(XA) \ge C'(1 + 2\sqrt{3})(\sqrt{m} + \sqrt{q})\sigma,$$

for some numerical constant $C' > 2\sqrt{2/7}$. *Then we have* $\mathbb{P}\{\widehat{k}_1 = r\} \to 1$ *for any* $N/2 \le r \le N$, *as* $N = q \wedge m \to \infty$.

In order to be able to select among ranks from $N/2$ to N in the first step, (3.11) requires $nm/\lambda_0 \ge (1+\delta)N/\delta$. However, Proposition 16 relaxes this to $nm/\lambda_0 \ge 3N/4$ from Proposition 16.

5. Extension to heavy tailed error distributions. Most results in this paper are finite sample results and apply to any matrix E. Only Corollary 5, Theorem 11 and the results in Section 4 require Gaussian errors. They appeal to precise concentration inequalities of $d_1(PE)$ around $\sqrt{m} + \sqrt{q}$, making use of the fact that $d_1(PE) = d_1(\Lambda U^T E)$ based on the eigendecomposition of $P = U\Lambda U^T$, and the fact that $\Lambda U^T E$ in turn is again Gaussian. In general, if E has independent entries, then the transformations PE or $\Lambda U^T E$ no longer have independent entries, although their columns remain independent. Regardless, our simulations reported in Sections 6.7 and 6.8 support our conjecture that our iterative method is flexible and our results continue to hold for general distributions, such as t-distributions with 6 degrees of freedom, for independent errors E_{ij}. For some important special cases, we are able to formally allow for errors with finite fourth moments only.

5.1. *Heavy tailed errors distributions with $n/q \to 1$.* We first consider the case $n/q \to 1$, which is likely to occur in high-dimensional settings ($p \gg n$). The following theorem guarantees the rank recovery via the GRS procedure for errors with heavy tailed distributions.

THEOREM 17. *Let $\lambda > 2(\sqrt{m} + \sqrt{q})^2$. Assume that the entries of E are i.i.d. random variables with mean zero and finite fourth moments. Furthermore, assume $n/q \to 1$, r and δ satisfy (3.10) and*

$$d_s(XA) \geq C\sigma(\sqrt{m} + \sqrt{q}), \tag{5.1}$$

for some $s \leq r$ and some numerical constant $C > 1 + \sqrt{2(1+\delta)}$. Then we have $\mathbb{P}\{s \leq \widehat{k} \leq r\} \to 1$ as $n \vee m \to \infty$, where $\widehat{k}$ is selected from (1.3).

In particular, if (5.1) holds for $s = r$, then $\widehat{k}$ consistently estimates r.

For a special case, that of skinny matrices XA, that is, $m = O(n^\alpha)$ or $n = O(m^\alpha)$ for some $0 \leq \alpha < 1$, we propose the following *Simplified Self-Tuning Rank Selection* (SSTRS) procedure. Given any $\varepsilon \in (0, 1)$, we set

$$\widehat{\lambda}_0 := 2(1+\varepsilon)(m \vee q), \qquad \widehat{k}_0 := \operatorname*{arg\,min}_{0 \leq k \leq \widehat{K}_0} \frac{\|Y - (Y)_k\|^2}{nm - \widehat{\lambda}_0 k} \tag{5.2}$$

as starting values, and if $\widehat{k}_0 \geq 1$, for $t \geq 0$, we update

$$\begin{aligned} \widehat{\lambda}_{t+1} &:= \frac{nm}{(1-\varepsilon)[(m \wedge q)/2 - \widehat{k}_t]_+ + \widehat{k}_t}, \\ \widehat{k}_{t+1} &:= \operatorname*{arg\,min}_{\widehat{k}_t \leq k \leq \widehat{K}_{t+1}} \frac{\|Y - (Y)_k\|^2}{nm - \widehat{\lambda}_{t+1} k}, \end{aligned} \tag{5.3}$$

where $[x]_+ := \max\{x, 0\}$ and $\widehat{K}_t := (nm/\widehat{\lambda}_t) \wedge q \wedge m$ for $t = 0, 1, \ldots$. The procedure stops when $\widehat{k}_t = \widehat{k}_{t+1}$ and we have the following result.

THEOREM 18. *Assume E_{ij} are i.i.d. random variables with mean zero and finite fourth moments. Suppose that $n/q \to 1$ and either $m = O(n^\alpha)$ or $n = O(m^\alpha)$ for some $\alpha \in [0, 1)$. Let $\{k_t\}_{t=0}^T$ be defined as Theorem 14. Define λ_0 as (5.2) and λ_{t+1} obtained from (5.3) by using k_t in lieu of $\widehat{k}_t$, for $t \geq 0$. Assume r and δ satisfy (3.10) for λ_0.*

Then, on the event

$$d_{k_t}(A) \geq C\sigma\sqrt{\lambda_t}, \qquad t = 0, \ldots, T \tag{5.4}$$

for some numerical constant $C > 1 + \sqrt{2(1+\delta)}$, there exists $0 \leq T' \leq T$ such that $\widehat{k}_{T'} = r$, with probability tending to 1, as $n \vee m \to \infty$. Here, $\widehat{k}_0 \leq \widehat{k}_1 \leq \cdots \leq \widehat{k}_{T'}$ are given in (5.2) and (5.3).

REMARK. As mentioned earlier, the entries of PE no longer inherit the independence from E when the distribution of the independent entries E_{ij} is not Gaussian. Nevertheless, by exploiting the independence of columns of PE, Theorem 5.39 in Vershynin (2012) shows that $d_1(PE) \le C_E\sqrt{q} + \sqrt{m}$ with high probability, provided the entries of E are independent sub-Gaussian random variables with unit variance. The constant C_E above unfortunately involves the unknown sub-Gaussian norm, which differs from σ^2. However, provided $q = o(m)$ and E has i.i.d. sub-Gaussian entries, we simply have $d_1(PE) \le (1 + o(1))\sqrt{m}$. Hence, for this case, it should be clear that our STRS procedure based on (4.2)–(4.5) can be directly applied with statistical guarantees stated in Section 4.

5.2. *A special model*: $Y = A + E$. We emphasize that our procedure can be applied to the important special model $Y = A + E$ where the entries of E are i.i.d. random variables with mean zero and finite fourth moments. The following results guarantee that our procedure can consistently estimate the rank of A. They are essentially the same statements as Theorems 17 and 18 for the case $Y = XA + E$, but this time *without* the disclaimer $n/q \to 1$.

THEOREM 19. *Assume the entries of $E \in \mathbb{R}^{n\times m}$ are i.i.d. random variables with mean zero and finite fourth moments. Assume further that r and δ satisfy* (3.10) *and*

$$d_s(A) \ge C\sigma(\sqrt{n} + \sqrt{m}) \tag{5.5}$$

for some $s \le r$ and some numerical constant $C > 1 + \sqrt{2(1+\delta)}$. Then $\mathbb{P}\{s \le \widehat{k} \le r\} \to 1$ as $m \vee n \to \infty$, where $\widehat{k}$ is selected from (5.2) *by using $\widehat{\lambda}_0 = \lambda > 2(\sqrt{n} + \sqrt{m})^2$.*

In particular, if (5.5) *holds for $s = r$, then $\widehat{k}$ consistently estimates r.*

In particular, when A is skinny, that is, $m = O(n^\alpha)$ or $n = O(m^\alpha)$ for some $0 \le \alpha < 1$, our newly proposed SSTRS in (5.2)–(5.3) maintains the rank consistency for this model.

THEOREM 20. *Let E_{ij} be i.i.d. random variables with mean zero and finite fourth moments, $m = O(n^\alpha)$ or $n = O(m^\alpha)$ for some $\alpha \in [0, 1)$, and assume r and δ satisfy* (3.10) *for λ_0 given in* (5.2). *Let $\{k_t\}_{t=0}^T$ be defined as Theorem 14 and λ_{t+1} obtained from* (5.3) *by using k_t in lieu of $\widehat{k}_t$, for $t \ge 0$. Then, on the event*

$$d_{k_t}(A) \ge C\sigma\sqrt{\lambda_t}, \qquad t = 0, \dots, T \tag{5.6}$$

for some numerical constant $C > 1 + \sqrt{2(1+\delta)}$, there exists $0 \le T' \le T$ such that $\widehat{k}_{T'} = r$, with probability tending to 1, as $m \vee n \to \infty$. Here, $\widehat{k}_0 \le \widehat{k}_1 \le \cdots \le \widehat{k}_{T'}$ are defined in (5.2) *and* (5.3).

6. Empirical study. The simulations in Sections 6.4 and 6.5 compare the methods discussed in this paper with some existing methods. Sections 6.6–6.8 verify our results for the proposed method. Our conclusions are summarized in Section 6.9. In Section 6.10, we perform an additional simulation to check the tightness of signal-to-noise condition in (3.9). In the Supplementary Material (Bing and Wegkamp (2019)), more simulations compare STRS using Monte Carlo simulations with STRS using deterministic bounds in Appendix F.1. We also present an example in Appendix F.2 to show that the method proposed by Giraud (2011) fails to recover the rank. Finally, simulations in Appendix F.3 compare STRS with other competing methods in terms of the errors $\|X\widehat{A} - XA\|$, $\|\widehat{A} - A\|$ and the selected rank.

6.1. *Methods and notation.* We first introduce the methods in our simulation. Bunea, She and Wegkamp (2011) proposed the method in (1.1) to select the optimal rank by using $\mu = Cd_1^2(Z)\widetilde{\sigma}^2$, where Z has $q \times m$ i.i.d. $N(0,1)$ entries and $\widetilde{\sigma}^2$ in (1.2) is the unbiased estimator of σ^2. The leading constant $C > 1$ needs to be specified. A deterministic upper bound which could be used instead is $C(m+q)\widetilde{\sigma}^2$. Bunea, She and Wegkamp (2011) suggests to use $C = 2$ based on its overall performance. However, there is no reason for one particular choice of C being globally optimal, which was confirmed in our simulations. Another option for choosing the tuning parameter is to use k-fold cross-validation. However, there is no theoretical guarantee of the feasibility for cross-validation, especially if the rows $X_{i\cdot}$ of X are non-i.i.d. In contrast, our proposed procedures (with and without self-tuning) are completely devoid of choosing a tuning parameter and estimating σ^2.

NOTATION. We use BSW to denote the method proposed in Bunea, She and Wegkamp (2011). For those methods proposed in this paper, we denote by GRS, STRS and SSTRS the method without self-tuning in (1.3), the one with self-tuning from (4.2)–(4.5) and the simpler version also with self-tuning from (5.2)–(5.3), respectively. We further use BSW-C to denote BSW with specified leading constant C.

6.2. *Task description.* We divide the simulation study up into five parts. In the first part, we show that there is no optimal constant C for the BSW-C method which works for all r. In the second part, we compare the performance of STRS and BSW-C (for various choices of C). The third part demonstrates the improvement of STRS over GRS shown in Section 4, in terms of requiring a smaller signal-to-noise ratio (SNR) and enlarging the range of possible selected ranks. The fourth part verifies the performance of GRS, STRS and SSTRS for non-Gaussian errors corresponding to our results and settings of Section 5. The last part extends the fourth part and supports our conjecture that STRS continues to work for general heavy tailed distributions under general settings.

6.3. *Simulation setup*. In general, we consider three settings. The first setting is the more favorable one where the sample size n is larger than the number of covariates p. The other two are high dimensional, $n < p$, and hence more challenging, with the last setting focussing on the worst case scenario of n close to q for a moderate m.

Our experiments are inspired by those of Bunea, She and Wegkamp (2011). When $n \geq p$, the $n \times p$ design matrix X is generated by independently drawing n times p-dimensional Gaussian vectors with mean zero and covariance matrix specified by $\Sigma_{i,j} = \eta^{|i-j|}$ with $\eta \in (0, 1)$, for $i, j = 1, \ldots, p$. When $n < p$, we let $X = X_1 X_2 \Sigma^{1/2}$ where $X_1 \in \mathbb{R}^{n \times q}$ and $X_2 \in \mathbb{R}^{q \times p}$ have i.i.d. $N(0, 1)$ entries. The regression coefficient matrix A is given by $A = b_0 M_{p \times r} M_{r \times m}$ where the entries of M are i.i.d. $N(0, 1)$. As before, r denotes the rank of A and satisfies $r \leq q \wedge m$. Regarding the error matrix E, each entry is generated from $N(0, 1)$ except in Sections 6.7 and 6.8 where we use t_ν-distributions with ν degrees of freedom.

The difference with Bunea, She and Wegkamp (2011) lies in the way we vary the signal-to-noise-ratio (SNR) defined as $d_r(XA)/\mathbb{E}[d_1(PE)]$. Instead of using various combinations of η and b_0, we vary the SNR by generating A with different ranks. Specifically, for given η and b_0, we first generate X, and then, for each r in some specified range, we generate A of rank r. For each pair (X, A), we generate 200 error matrices E, calculate the SNR and record the rank recovery rate and the mean selected rank for various methods in the 200 replications.

6.4. *Experiment* 1. We compare the rank recovery of BSW-C for C in $\{0.7, 0.9, 1.1, 1.3, 1.5\}$ with STRS in both low- and high-dimensional settings. In the low-dimensional case, we consider $n = 150$, $m = 30$, $p = q = 20$, $r \in \{0, \ldots, 20\}$ and $\eta = 0.1$. For b_0, we choose from $\{0.15, 0.20, 0.25\}$ to illustrate, more clearly, the effect of r on the recovery rate. The high-dimensional setting has $n = 100$, $m = 30$, $p = 150$, $q = 20$, $\eta = 0.1$, $b_0 = \{0.03, 0.05, 0.07\}$ and $r \in \{0, \ldots, 20\}$. The rank recovery rate and mean selected ranks for both low- and high-dimensional cases are shown in Figures 2 and 3, respectively.

RESULT. Both figures demonstrate that BSW-C with a smaller C, say 0.7 or 0.9, performs better when the true rank r is large, in the sense of requiring a smaller SNR, but tends to overfit for small r. In contrast, BSW-C with a larger C does a better job in preventing overfitting for small r, but requires a larger SNR. The performance of BSW-C does not seem to depend on r as we separate the effect of SNR away from this phenomenon by varying b_0. This suggests that there is no optimal leading constant C for BSW-C to guarantee consistent rank estimation for all possible r. On the other hand, STRS performs globally better and more stable than BSW-C in all settings. It prevents overfitting for both small and large r and requires a smaller SNR than BSW-1.3 and BSW-1.5. Finally, the role of SNR for the rank recovery is striking. If it is too small (less than 0.8), we completely fail to recover the rank. This justifies the signal-to-noise condition in (3.9) and is explained by the (fast) exponential tail bounds in our main results.

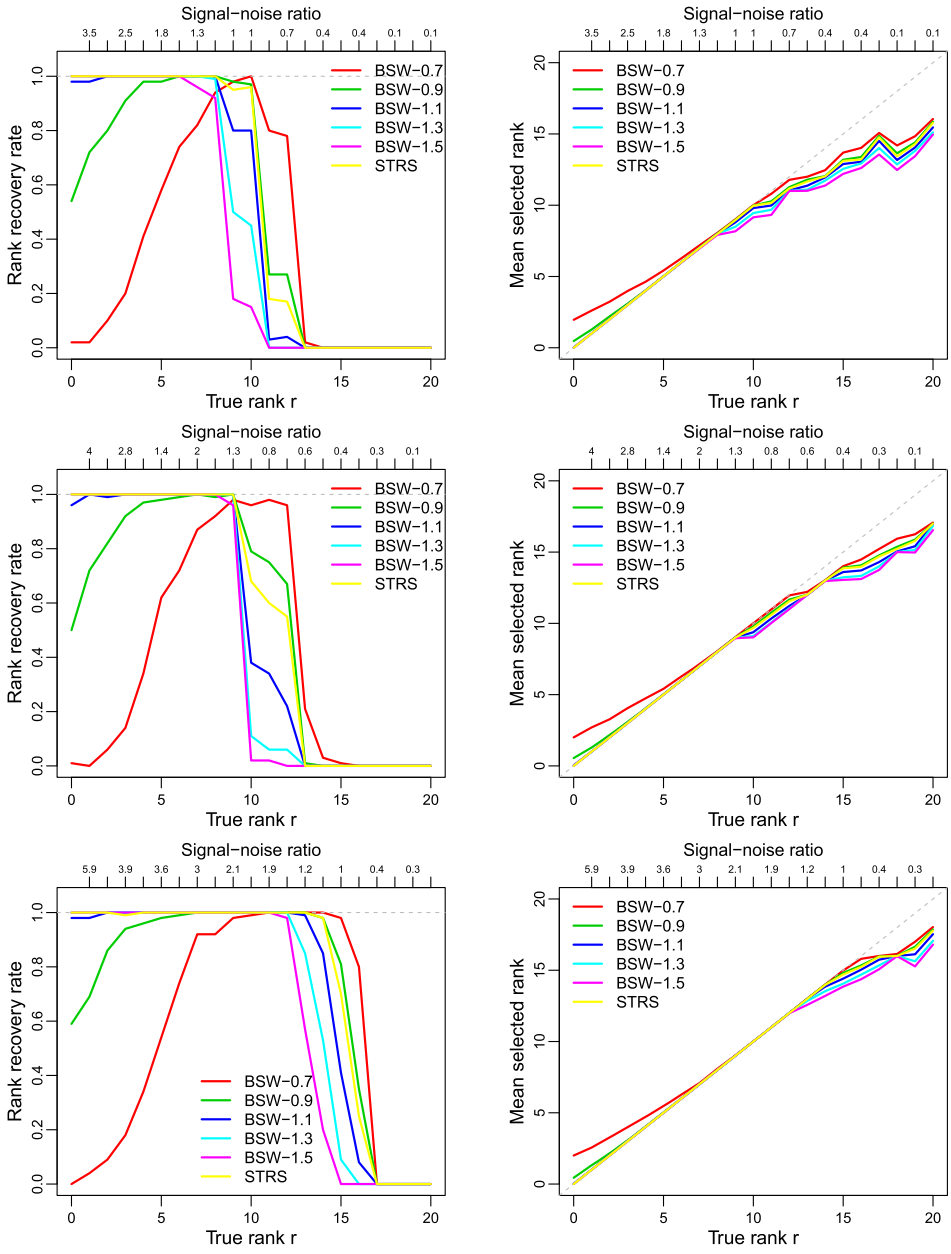

FIG. 2. *Comparison of BSW-C and STRS in the low-dimensional setting of Experiment* 1. *Here*, b_0 *is* 0.15 (*top*), 0.20 (*middle*) *and* 0.25 (*bottom*).

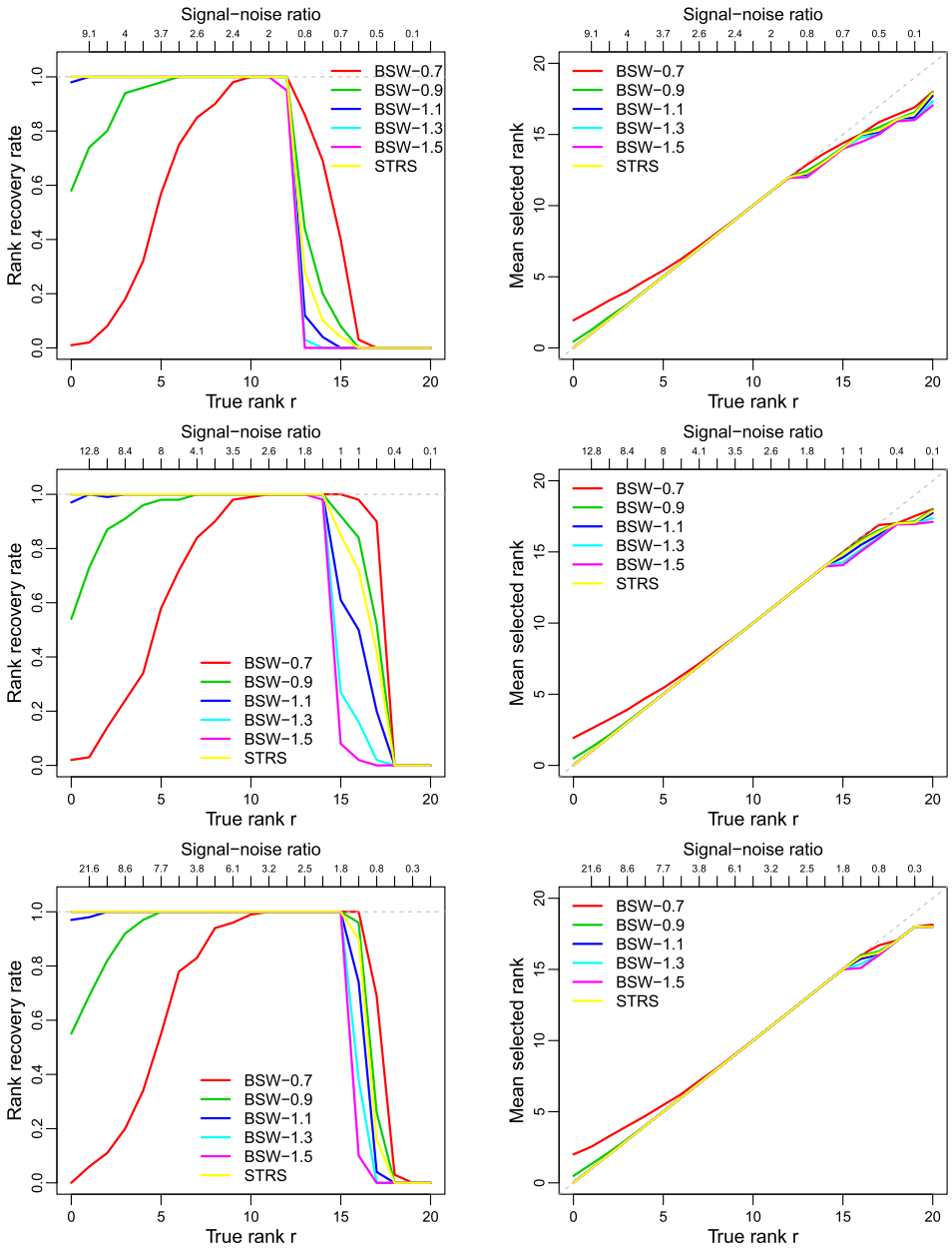

FIG. 3. *Comparison of BSW-C and STRS in the high-dimensional setting of Experiment* 1. *Here* b_0 *is* 0.03 (*top*), 0.05 (*middle*) *and* 0.07 (*bottom*).

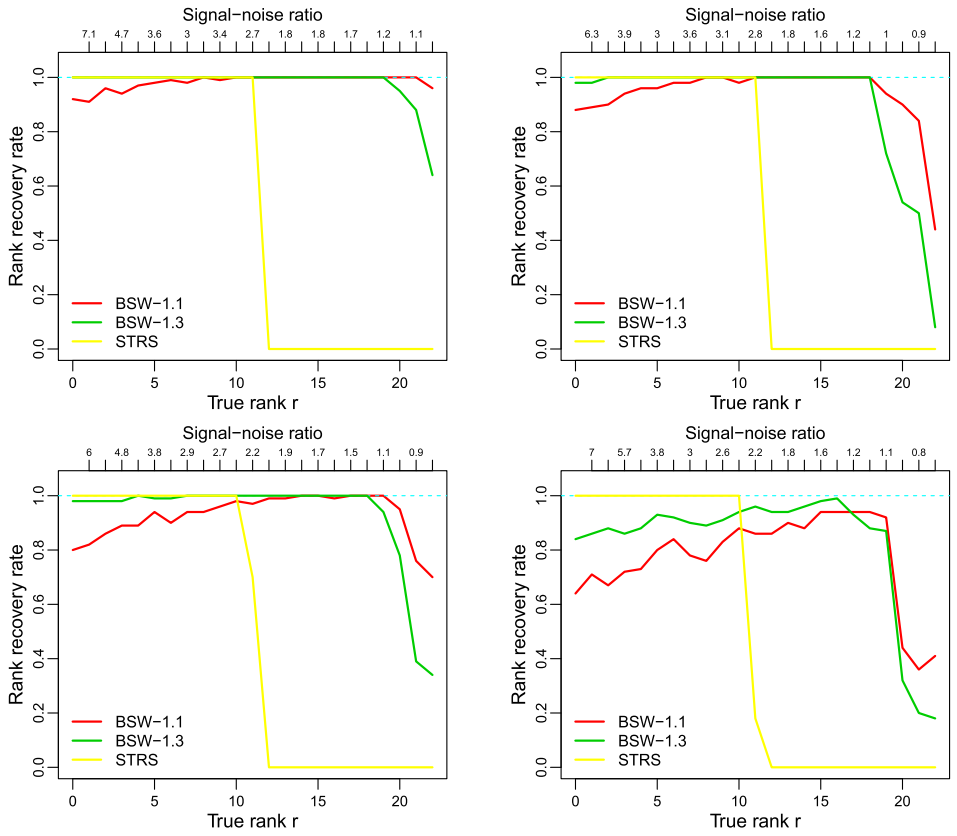

FIG. 4. *Plots of Experiment* 2 *on rank recovery rate for BSW*-1.1, *BSW*-1.3 *and STRS. The plots from left to right and top to bottom have* {143, 145, 147, 149} *for* q.

6.5. *Experiment* 2. Based on the results in Section 6.4, we compare BSW-1.1 and BSW-1.3 with STRS. Figures 2 and 3 show the advantage of STRS over BSW-C in both low- and high-dimensional settings when n is not too small compared to q. Here, we focus on the worst case scenario of $n \approx q$ and set $n = 150$, $m = 30$, $p = 200$, $\eta = 0.1$, $b_0 = 0.011$, $q \in \{143, 145, 147, 149\}$ and $r \in \{0, \ldots, 22\}$. The recovery rates are shown in Figure 4.

RESULT. We see that, for moderate m, BSW-C performs worse as n gets closer to q. Indeed, estimation of σ^2 is problematic for small values of $(n - q)m$. The same problem for choosing different C persists: BSW-1.1 requires a smaller SNR, but overfits more than BSW-1.3 at small r, while STRS performs perfectly as long as r lies in the allowable range (its largest recoverable rank is between 11 and 13, depending on the particular choice of q).

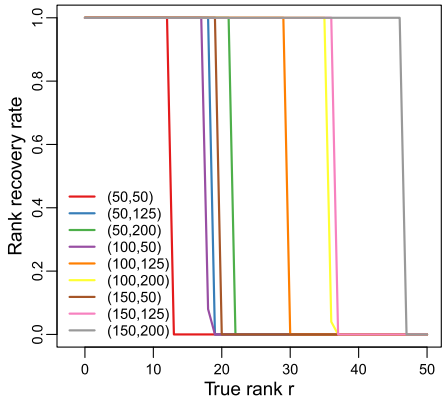

FIG. 5. *Rank recovery rate for STRS when $n = q$ in Experiment* 2.

In addition, we verify the feasibility of STRS when $n = q$. In this case, BSW is infeasible. We consider $p = 200$, $b_0 = 0.011$, $\eta = 0.1$, $n = q \in \{50, 100, 150\}$, $m \in \{50, 125, 200\}$ and $r \in \{0, \ldots, 50\}$. In each setting, the signal is large enough (SNR > 3) to eliminate the effect of the signal-to-noise ratio on rank recovery. Figure 5 shows that STRS recovers the rank for all combinations of n and m as long as r is within the recoverable range (which increases in m).

6.6. *Experiment* 3. Figure 6 demonstrates the advantages of STRS over GRS, stated in Theorem 15 and Proposition 16, in three settings. The first setting is the same low-dimensional scenario considered in Experiment 1 with $b_0 = 0.25$. The second setting uses the same high-dimensional setting considered in Experiment 1 with $b_0 = 0.07$. The third setting focuses on the case when $nm \le \lambda_0 N$, which incurs the rank constraint (3.10), and we set $n = 50$, $m = 50$, $p = 300$, $q = 30$, $\eta = 0.1$, $b_0 = 2$ and $r \in \{0, \ldots, 30\}$. In this setup, $K_{\lambda_0} = 7$.

RESULT. The top two figures in Figure 6 indicate that STRS requires a SNR of about 1, while GRS needs a SNR of about 2 for correct recovery. The bottom two plots in Figure 6 show that GRS fails to recover the rank r if $r > K_{\lambda_0}$, whereas STRS perfectly recovers all possible ranks. This confirms that when nm is not too small compared to $(\sqrt{m} + \sqrt{q})^2$, STRS can get rid of the rank constraint (3.10). (The tuning parameter λ_t in STRS reduces from 198 to 83 and 66 in the first two cases, respectively, and from 315 to 69 in the third case.) These findings confirm our theoretical results in Section 4.

6.7. *Experiment* 4. We verify the results of Section 5 by comparing the performance of GRS, STRS and SSTRS for both models $Y = XA + E$ and $Y = A + E$ with errors E_{ij} generated from a t_ν-distribution with various degrees of freedom ν.

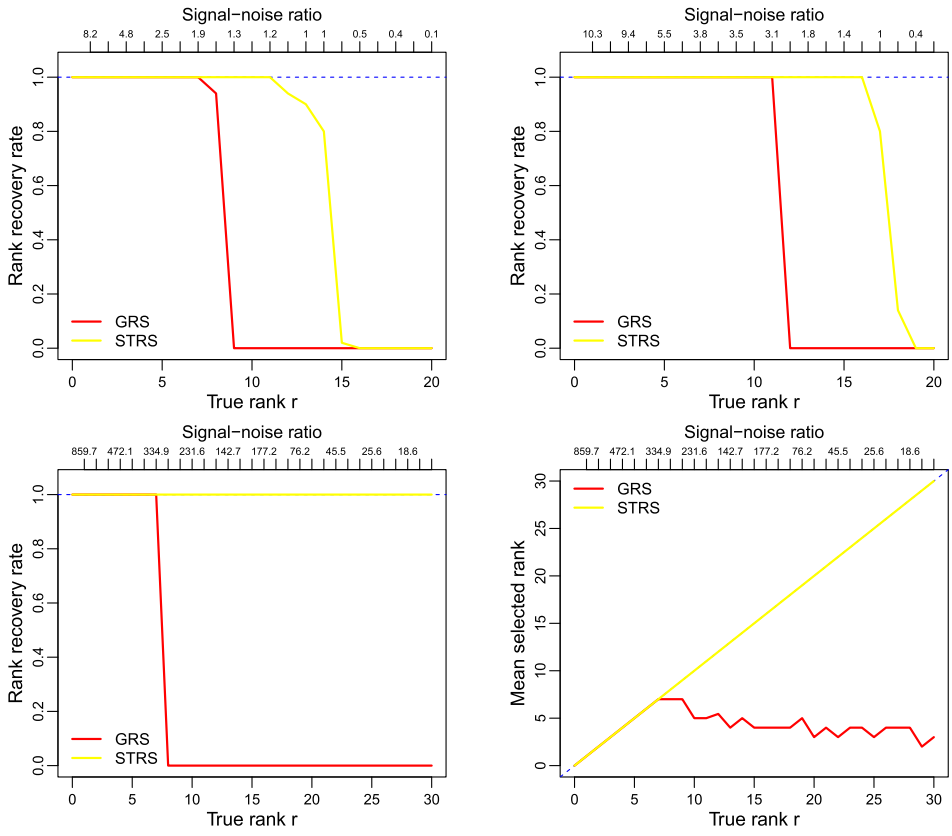

FIG. 6. *Comparison of GRS and STRS in Experiment* 3 *in the first setting* (*top left*), *second setting* (*top right*) *and third setting* (*bottom*).

For model $Y = XA + E$, we consider the case $n/q \to 1$ with different m. We set $\eta = 0.1$ and $r \in \{0, \ldots, 15\}$ for all settings. The first plot in Figure 7 depicts mean selected ranks of GRS and STRS when we further set $n = q = 150$, $m = 100$, $p = 250$ and $b_0 = 0.002$ and $\nu = 6$ (degrees of freedom of the t_ν distribution). We also verify the rank consistency of SSTRS by generating E_{ij} from t_ν-distributions with $\nu \in \{6, 8, 10\}$. The second row in Figure 7 depicts mean selected ranks of SSTRS and is based on $n = 300 \approx q = 280 \gg m = 50$, $p = 400$ and $b_0 = 0.0015$. The third row in Figure 7 shows the same quantities and is based on $n = 80$, $q = 60$, $p = 150$, $m = 400$ and $b_0 = 0.003$. We varied the closeness of n and q, but since the results did not change, we only report for one pair of n and q for each setting.

For model $Y = A + E$, we present two cases of skinny A when $n = O(m^\alpha)$ and $m = O(n^\alpha)$ for some $\alpha \in (0, 1)$. Specifically, we consider $n = 500$, $m = 80$ in the first setting and $n = 80$, $m = 500$ in the second one. We set $\eta = 0.1$, $b_0 = 0.25$, $r \in \{0, \ldots, 20\}$ and $\nu \in \{6, 8, 10\}$ in both cases. The mean selected ranks of SSTRS are plotted in the last row of Figure 7.

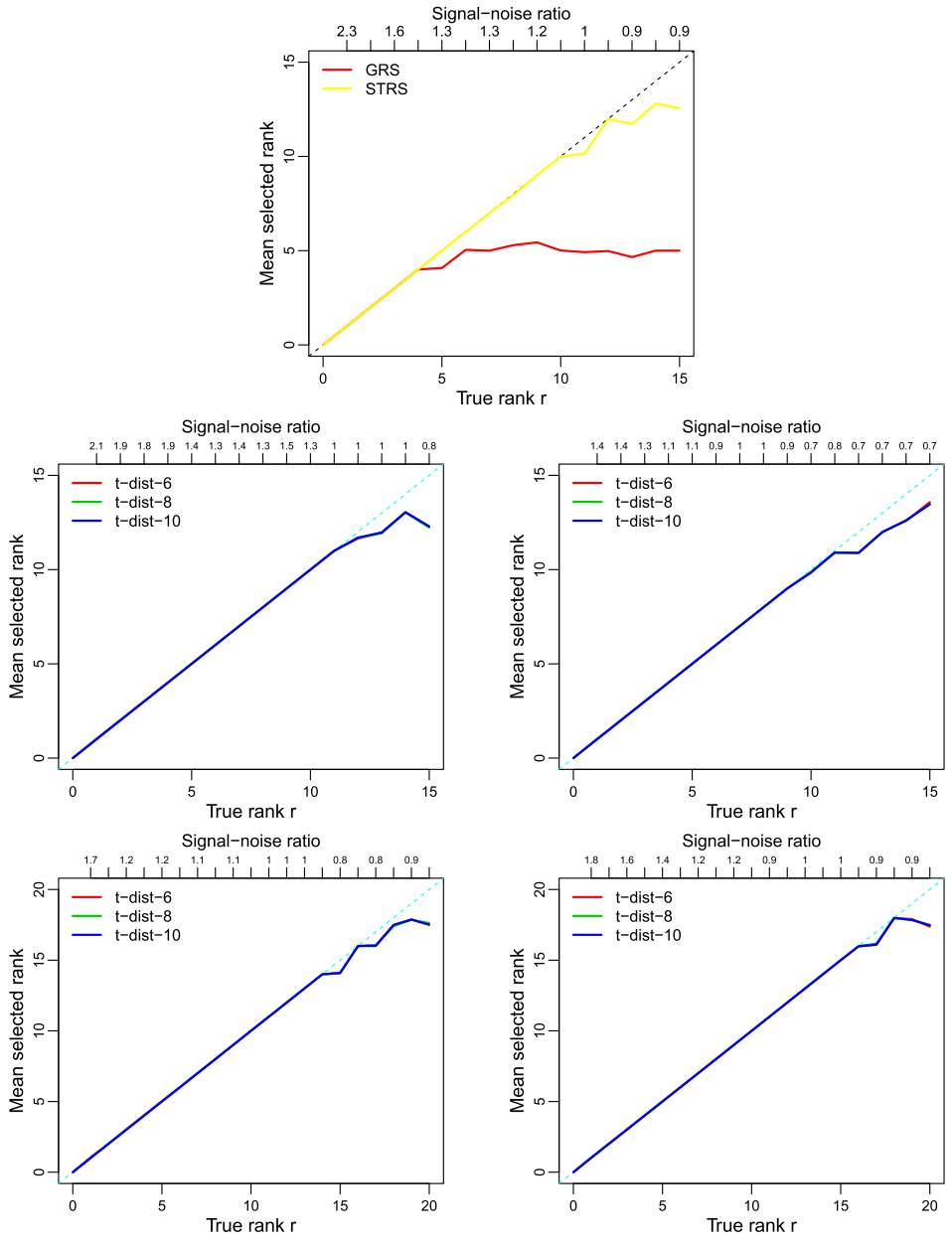

FIG. 7. *Plots of mean selected ranks related to Experiment* 4. *The first plot compares GRS and STRS in model* $Y = XA + E$. *The middle row evaluates SSTRS in model* $Y = XA + E$ *for various error distributions with* $n = 300$, $q = 280$, $m = 50$ (*left*) *and* $n = 80$, $q = 60$, $m = 400$ (*right*). *The bottom row plots mean selected ranks of SSTRS in model* $Y = A + E$ *with* $n = 500$, $m = 80$ (*left*) *and in* $m = 80$, $n = 500$ (*right*).

RESULT. In both models, all three procedures work perfectly for heavy tailed errors under very mild SNR, although the rank constraint (3.10) impacts GRS. In addition, STRS and SSTRS can handle a larger range of r under a milder SNR and their performance seems very stable under various error distributions.

6.8. *Experiment* 5. The stable performance of GRS, STRS and SSTRS leads us to make the following conjecture:

$$\mathbb{E}\big[d_j(PE)\big] \approx \mathbb{E}\big[d_j(Z)\big] \qquad \text{for all } j = 1, \ldots, q \wedge m, \tag{6.1}$$

where $Z \in \mathbb{R}^{q\times m}$ has i.i.d. $N(0,1)$, P is the projection matrix based on X with $\operatorname{rank}(P) = q$ and entries of $E \in \mathbb{R}^{n\times m}$ are i.i.d. mean zero random variables with $\mathbb{E}[E_{ij}^2] = 1$ and $\mathbb{E}[E_{ij}^4] < \infty$. The result is striking since the projection P destroys the independence of E_{ij}, hence one would not necessarily expect the Bai–Yin law (Bai and Yin (1993)) continue to hold for PE which only has independent columns. Proving (6.1) is beyond the scope of the current paper and we leave it for future research. Instead, we verify this conjecture in simulations for two cases: (1) $n = 150$, $p = 250$, $q = 50$, $m = 50$; (2) $n = 50$, $p = 40$, $q = 40$, $m = 150$. In both cases, $\eta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and we generate E from t_ν-distributions with degrees of freedom $\nu \in \{5, 8, 12\}$. For each setting, we generate X and P for a given η, and we generate 100 pairs of matrices E and Z. Averaged ratios of $d_j(PE)/d_j(Z)$ are calculated for each j and Figure 8 shows that the ratios of $d_j(PE)/d_j(Z)$ are highly concentrated around 1. We only report the case of $\eta = 0.9$ as the other cases gave essentially the same picture.

In light of this, we further conjecture that our procedures work in general settings with heavy tailed error distributions. We consider both low- and high-dimensional settings to verify this claim. The low-dimensional setting considers $n = 150$, $p = q = m = 30$ and $b_0 = 0.15$ and the high-dimensional setting considers $n = 100$, $p = 150$, $q = m = 30$ and $b_0 = 0.015$. We generate E from t_ν-distribution with $\nu \in \{6, 8, 10\}$ and we set $\eta = 0.1$ and $r \in \{0, \ldots, 20\}$ in both

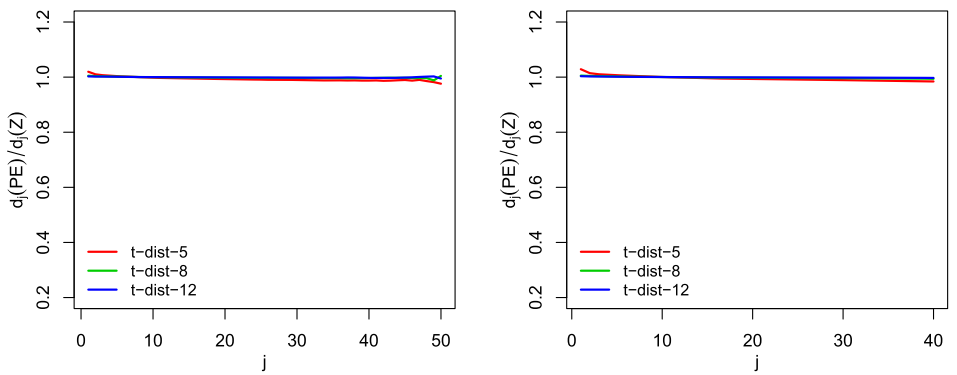

FIG. 8. *Panel of* $\mathbb{E}[d_j(PE)]/\mathbb{E}[d_j(Z)]$ *in Experiment* 5 *with* $n = 150$, $p = 250$, $m = q = 50$ (*left*) *and* $n = 50$, $p = q = 40$, $m = 150$ (*right*).

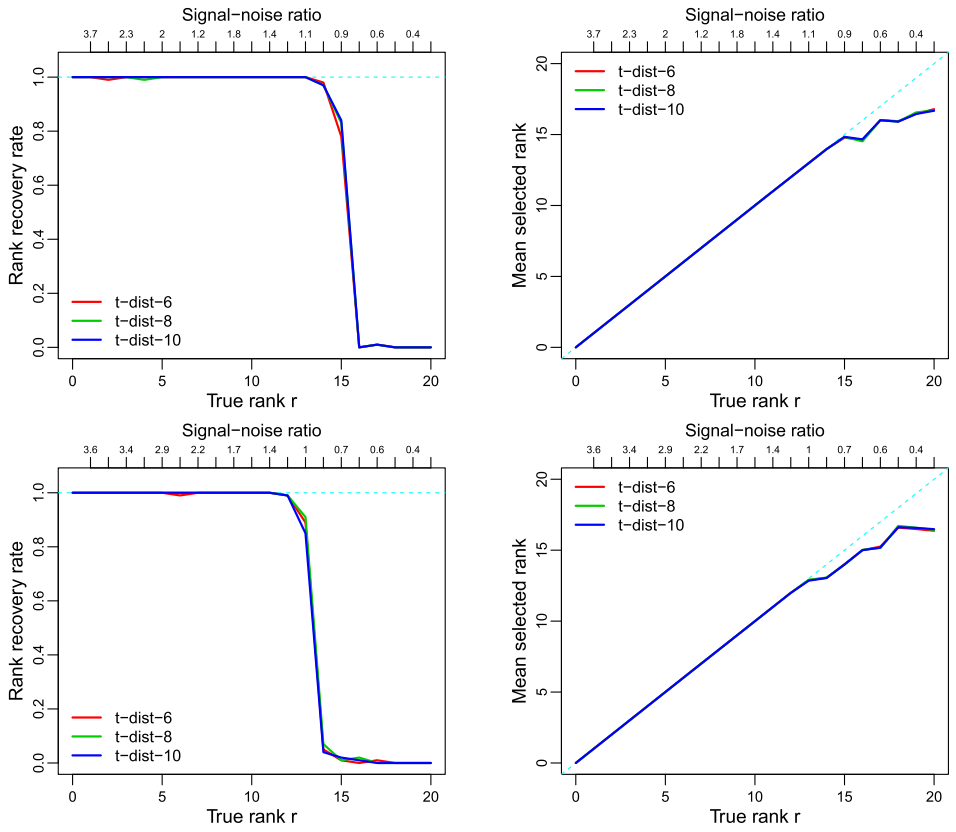

FIG. 9. *Performance of STRS for heavy tails in Experiment* 5 *in the low-dimensional* (*top*) *and the high-dimensional setting* (*bottom*).

cases. The plots in Figure 9 show that STRS consistently estimates the rank in both settings under a very mild SNR ratio and its performance is quite stable for different heavy tailed t-distributions.

6.9. *Conclusions of the simulation studies.*

- In general, STRS outperforms BSW-C in both low-dimensional and high-dimensional settings. The performance of BSW-C is influenced by the true rank r and there is no globally optimal tuning parameter C for BSW-C. STRS is stable in general as long as the true rank r lies in its allowable range.
- In the most challenging setting of Experiment 2, when $n \approx q$ and estimation of σ^2 is problematic, the advantage of STRS over BSW-1.1 and BSW-1.3 becomes more prominent. If $n = q$, BSW-C is no longer feasible, while STRS only fails in the rare situation when nm is small compared to $m + q$ and r is large. Of course, reduced rank regression only makes sense for relatively small r.

- Experiment 3 verifies that STRS has clear advantages over GRS. It requires a smaller signal-to-noise ratio and allows for larger values of r, which confirms our theoretical result in Section 4.
- Experiment 4 confirms our results in Section 5, that our procedures (GRS, STRS, SSTRS) continue to consistently estimate the true rank for heavy tailed distributions in certain settings, considered in Section 5. Moreover, Experiment 5 confirms our conjecture that STRS works in more general settings, even if the errors are generated from heavy tailed distributions.

6.10. *Tightness check of signal-to-noise condition.* Akin to the discussion in Bunea, She and Wegkamp ((2011), Section 4.2, page 1303), we can empirically verify the tightness of the signal-to-noise condition in (3.9). Specifically, from (2.7), we have

$$\{\widehat{k} \neq r\} = \{d_{r+1}(PY) \geq \sqrt{\lambda}\widehat{\sigma}_r\} \cup \{d_r(PY) \leq \sqrt{\lambda}\widehat{\sigma}_r\}. \tag{6.2}$$

By using identity (6.2) and Weyl's inequality, we observe that

$$\mathbb{P}\{\widehat{k} \neq r\} \geq \mathbb{P}\{d_r(XA) + d_1(PE) \leq \sqrt{\lambda}\widehat{\sigma}_r\}.$$

Hence we conclude that $\mathbb{P}\{d_r(XA) \leq \sqrt{\lambda}\widehat{\sigma}_r - d_1(PE)\} > 0$ implies $\mathbb{P}\{\widehat{k} = r\} < 1$. This suggests $d_r(XA)$ cannot be smaller than $\sqrt{\lambda}\widehat{\sigma}_r - d_1(PE)$. To empirically verify this conjecture, we generate different pairs of (X, A) through changing b_0, η, n, m, p, q and r. For each pair of (X, A), we record the rth largest singular value of XA as $d_r(XA)$ and we search along a grid of λ to find the largest λ such that minimizing (1.3) recovers the true rank (recall that $\sqrt{\lambda}\widehat{\sigma}_r$ is increasing in λ). Finally, we plot $\lambda\widehat{\sigma}_r$ and $\sqrt{\lambda}\widehat{\sigma}_r - d_1(PE)$ against $d_r(XA)$ for all pairs of (X, A) in Figure 10. This plot collaborates our conjecture that the signal-to-noise condition in (3.9) is tight.

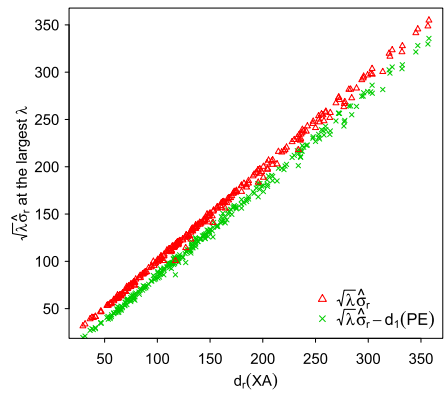

FIG. 10. *Plot of $\sqrt{\lambda}\widehat{\sigma}_r$ and $\sqrt{\lambda}\widehat{\sigma}_r - d_1(PE)$ versus $d_r(XA)$ for each pair of (X, A). The value for λ is the largest one (on a grid) that correctly found the true rank.*

Acknowledgments. The authors thank the Editor, Associate Editor and two referees for constructive remarks.

SUPPLEMENTARY MATERIAL

Supplement to "Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models" (DOI: 10.1214/18-AOS1774SUPP; .pdf). The supplementary document includes the oracle inequality for the fit, additional simulation results and all proofs.

REFERENCES

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat.* **22** 327–351. MR0042664

ANDERSON, T. W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *Ann. Statist.* **27** 1141–1154. MR1740118

ANDERSON, T. W. (2002). Specification and misspecification in reduced rank regression. *Sankhyā Ser. A* **64** 193–205. MR1981753

BAI, Z. D. and YIN, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294. MR1235416

BING, X. and WEGKAMP, M. H (2019). Supplement to "Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models." DOI:10.1214/18-AOS1774SUPP.

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

BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355

BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.* **40** 2359–2388. MR3097606

ECKART, C. and YOUNG, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

GIRAUD, C. (2011). Low rank multivariate regression. *Electron. J. Stat.* **5** 775–799. MR2824816

GIRAUD, C. (2015). *Introduction to High-Dimensional Statistics*. *Monographs on Statistics and Applied Probability* **139**. CRC Press, Boca Raton, FL. MR3307991

IZENMAN, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** 248–264. MR0373179

IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques*: *Regression*, *Classification*, *and Manifold Learning*. *Springer Texts in Statistics*. Springer, New York. MR2445017

JOHNSTONE, I. M. (2001). Chi-square oracle inequalities. In *State of the Art in Probability and Statistics*: *Festschrift for Willem R. van Zwet* (M. de Gunst, C. Klaasen and A. van der Vaart, eds.) 399–418. IMS, Beachwood, OH. MR1836572

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

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

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839

RAO, C. R. (1978). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Multivariate Analysis V* 3–22. North-Holland, Amsterdam.

REINSEL, G. C. and VELU, R. P. (1998). *Multivariate Reduced-Rank Regression*: *Theory and Applications*. *Lecture Notes in Statistics* **136**. Springer, New York. MR1719704

ROBINSON, P. M. (1973). Generalized canonical analysis for time series. *J. Multivariate Anal.* **3** 141–160. MR0326959

ROBINSON, P. M. (1974). Identification, estimation and large-sample theory for regressions containing unobservable variables. *Internat. Econom. Rev.* **15** 680–692. MR0356376

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

RUDELSON, M. and VERSHYNIN, R. (2010). Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III* 1576–1602. Hindustan Book Agency, New Delhi. MR2827856

SCHMIDT, E. (1907). Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. *Math. Ann.* **63** 433–476.

STEWART, G. W. (1993). On the early history of the singular value decomposition. *SIAM Rev.* **35** 551–566. MR1247916

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.* **71** 441–479. MR1511670

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
301D MALOTT HALL
ITHACA, NEW YORK 14853-3801
USA
E-MAIL: xb43@cornell.edu

DEPARTMENT OF MATHEMATICS
AND DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
432 MALOTT HALL
ITHACA, NEW YORK 14853-3801
USA
E-MAIL: marten.wegkamp@cornell.edu