Electronic Journal of Statistics
Vol. 14 (2020) 1348–1385
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1681

Computing the degrees of freedom of rank-regularized estimators and cousins

Rahul Mazumder

MIT Sloan School of Management,
Operations Research Center and Center for Statistics,
Massachusetts Institute of Technology, Cambridge, MA 02142
e-mail: rahulmaz@mit.edu

Haolei Weng

Department of Statistics and Probability,
Michigan State University, East Lansing, MI 48824
e-mail: wenghaol@msu.edu

Abstract: Estimating a low rank matrix from its linear measurements is a problem of central importance in contemporary statistical analysis. The choice of tuning parameters for estimators remains an important challenge from a theoretical and practical perspective. To this end, Stein's Unbiased Risk Estimate (SURE) framework provides a well-grounded statistical framework for degrees of freedom estimation. In this paper, we use the SURE framework to obtain degrees of freedom estimates for a general class of spectral regularized matrix estimators—our results generalize beyond the class of estimators that have been studied thus far. To this end, we use a result due to Shapiro (2002) pertaining to the differentiability of symmetric matrix valued functions, developed in the context of semidefinite optimization algorithms. We rigorously verify the applicability of Stein's Lemma towards the derivation of degrees of freedom estimates; and also present new techniques based on Gaussian convolution to estimate the degrees of freedom of a class of spectral estimators, for which Stein's Lemma does not directly apply.

MSC 2010 subject classifications: 62H12.
Keywords and phrases: Degrees of freedom, divergence, low rank, matrix valued function, regularization, spectral function, SURE.

Received September 2019.

Contents

1. Introduction

Consider the basic sequence model setup with

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \quad \mathrm{Cov}(\boldsymbol{\epsilon}) = \tau^2 \mathbf{I}, \quad \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0},$$

where, $\mathbf{I}$ is the identity matrix; and we observe $\mathbf{y} \in \mathbb{R}^n$, a noisy version of the unknown signal $\boldsymbol{\mu} \in \mathbb{R}^n$. Let $\hat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$. The accuracy of $\hat{\boldsymbol{\mu}}$ as an estimator for $\boldsymbol{\mu}$ is often quantified via the expected mean squared error (MSE) which admits the following decomposition [9]

$$R \triangleq \mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 = -\tau^2 n + \mathbb{E}\|\hat{\boldsymbol{\mu}} - \mathbf{y}\|_2^2 + 2\sum_{i=1}^{n} \mathrm{Cov}(\hat{\mu}_i, y_i), \tag{1.1}$$

where $\|\cdot\|_2$ is the usual ℓ_2 norm, and the subscript i indicates the ith component of a vector. The covariance term appearing in (1.1) measures the complexity of the estimator $\hat{\boldsymbol{\mu}}$ and is related to the well known degrees of freedom (df) of an estimator [29, 9]:

$$df(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \mathrm{Cov}(\hat{\mu}_i, y_i)/\tau^2. \tag{1.2}$$

The decomposition (1.1) suggests an unbiased estimator $\widehat{df}\,(\hat{\boldsymbol{\mu}})$ for $df(\hat{\boldsymbol{\mu}})$ that leads to an unbiased estimate for R:

$$\widehat{R} = -\tau^2 n + \|\hat{\boldsymbol{\mu}} - \mathbf{y}\|_2^2 + 2\tau^2 \cdot \widehat{df}\,(\hat{\boldsymbol{\mu}}). \tag{1.3}$$

We can then use $\widehat{R}$ to choose between different estimators. Hence the degrees of freedom plays an important role in model assessment and selection. Consider the example of multiple linear regression, where $\boldsymbol{\mu} = X\boldsymbol{\beta}$ with design matrix $X \in \mathbb{R}^{n\times p}$ and regression coefficient $\boldsymbol{\beta} \in \mathbb{R}^p$. In the case when $n > p$ and X is of full rank, the *df* of the least square estimates equals p, i.e., the number of parameters in the model. This fact combined with (1.3) leads to the well known Mallows's C_p criterion [20]. For estimators $\hat{\boldsymbol{\mu}}$ that are a linear functional of $\mathbf{y}$ (arising via ridge regression, for example), the *df* can be computed by looking at the trace of the smoother matrix [12]. However, for estimators that are nonlinear functionals of $\mathbf{y}$, the computation of *df* becomes much more challenging. [29, 9] derive an alternate expression of *df* for the Gaussian sequence model $\mathbf{y} \sim N(\boldsymbol{\mu}, \tau^2\mathbf{I})$ when $\hat{\boldsymbol{\mu}}$ is weakly differentiable[1] with respect to $\mathbf{y}$. In this case, the degrees of freedom of $\hat{\boldsymbol{\mu}}$ is given by the well-known Stein's Lemma:

$$\text{(Stein's Lemma)} \qquad df(\hat{\boldsymbol{\mu}}) = \mathbb{E}\left(\sum_{i=1}^{n} \partial\hat{\mu}_i/\partial y_i\right) \tag{1.4}$$

which suggests an unbiased estimate for R, termed Stein's Unbiased Risk Estimate (SURE):

$$\widehat{R} = -\tau^2 n + \|\hat{\boldsymbol{\mu}} - \mathbf{y}\|_2^2 + 2\tau^2 \cdot \sum_{i=1}^{n} \partial\hat{\mu}_i/\partial y_i.$$

The SURE framework has been successfully utilized in different statistical problems. For instance, [5] derived the *df* of soft thresholding in a wavelet shrinkage procedure. [42, 34] studied the *df* of lasso and generalized lasso fit. [21, 33] obtained the *df* of best subset selection under the linear regression model with orthogonal design.

The above framework also applies to matrix estimation — here, data is of the form $y_{ij} = \mu_{ij} + \epsilon_{ij}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. The general problem of low rank matrix estimation has been widely studied in the statistical community in the context of multivariate linear regression [1, 16, 39] and matrix completion [3, 22], among others. There has been nice recent work on using SURE theory to derive the *df* of low rank matrix estimators – but the problem becomes quite challenging as one needs to deal with the differentiability properties of nonlinear functions of the spectrum and singular vectors of a matrix. Candès et al. [4] obtained the analytic expression of the divergence[2] $\sum_{ij} \partial\hat{\mu}_{ij}/\partial y_{ij}$ for a singular value thresholding estimator – they also rigorously verified sufficient conditions under which Stein's Lemma holds. [25, 38] derived expressions for the divergence of certain reduced rank and nuclear norm penalized estimators; but they do not appear to formally verify if the regularity conditions sufficient for Stein's Lemma to hold, are satisfied. To sum up, the challenge for deriving the *df* of matrix estimators is three-fold. Firstly, it may be challenging to

[1] There are additional mild integrability conditions about $\hat{\boldsymbol{\mu}}$. Please refer to Appendix 7.9 for details.

[2] See the formal definition in Section 1.1.

verify the regularity conditions required for (1.4) to hold. A direct plug-in of formula (1.4) may lead to inaccurate *df* calculation[3]. Secondly, even when formula (1.4) is available, it might be difficult to derive an analytical expression of $\sum_{ij} \partial \hat{\mu}_{ij}/\partial y_{ij}$, especially for matrix estimators that depend on the singular vectors/values of the observed matrix in a non-linear way. Thirdly, there are estimators for which Stein's Lemma is not readily applicable – in these cases, new techniques may be necessary to derive *df* estimates. Thusly motivated, in this paper, we aim to present a systematic study of two generic low rank matrix estimators, namely spectral regularized and rank constrained estimators—this includes, but is not limited to, all estimators studied in the three aforementioned works. Our contributions are summarized as:

(i) We propose a framework to derive the analytic formula of $\sum_{ij} \partial \hat{\mu}_{ij}/\partial y_{ij}$ for general matrix estimators, by appealing to some nice (but seemingly underutilized) results pertaining to differentiability of symmetric matrix valued functions due to Shapiro [28]—these results were derived in the context of semidefinite optimization algorithms. The expressions for the *df* of several estimators are thus shown to follow as special cases.
(ii) For several matrix estimators where Stein's Lemma is not directly applicable, our derivation of the *df* relies on using ideas from Gaussian convolution along with subtle limiting arguments that utilize the eigenvalue distribution of a real-valued central Wishart matrix. The techniques proposed in this paper may apply to a wider class of estimators, beyond what is studied herein.
(iii) Our analysis covers a much wider range of low rank matrix estimators than what has been studied before, and we present a unified framework to address these problems.

The remainder of the paper is organized as follows. We introduce the main theorem for calculating the divergence of matrix estimators in Section 2. Sections 3 and 4 consist of multiple applications of the main theorem in deriving the degrees of freedom for various low rank matrix estimators. Numerical experiments are performed in Section 5 to validate the derived *df* formulas. We present our concluding remarks in Section 6. All of our proofs and related technical material are relegated to the appendix.

1.1. Notations

For a vector $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathbb{R}^n$, we use the notation $\mathrm{diag}(\boldsymbol{a})$ to denote the $n \times n$ diagonal matrix with ith diagonal entry being a_i. For a real matrix $Y \in \mathbb{R}^{m \times n}$ (we assume, without loss of generality, $m \geq n$ throughout the paper), let its transpose be Y' and its reduced singular value decomposition be $Y = U\mathrm{diag}(\boldsymbol{\sigma})V'$, where $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n), V = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n), \boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$

[3]For example, in the best subset selection procedure in linear regression, the formula does not hold and the *df* estimate is not the number of nonzero regressors.

and $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. We denote the Frobenius norm of Y by $\|Y\|_F$. Unless otherwise stated, we use $Y = U\text{diag}(\boldsymbol{\sigma})V'$ to represent the reduced singular value decomposition (SVD). Y is called simple if it has no repeated singular values. For a real valued function $f : \mathbb{R}^+ \to \mathbb{R}$, define the associated matrix valued spectral function $S(\cdot; f) : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ as $S(Y; f) = U\text{diag}(f(\boldsymbol{\sigma}))V'$ where $f(\boldsymbol{\sigma}) = (f(\sigma_1), \ldots, f(\sigma_n))$. A function f is said to be directionally differentiable at x if the directional derivative

$$f'(x; h) \triangleq \lim_{t\downarrow 0} \frac{f(x + th) - f(x)}{t} \tag{1.5}$$

exists for any h. Denote the divergence of $S(Y; f)$ by

$$\nabla \cdot S(Y; f) \triangleq \sum_{ij} \partial[S(Y; f)]_{ij}/\partial Y_{ij},$$

where $[S(Y; f)]_{ij}$ is the (i, j)th element of $S(Y; f)$. When we mention regularity conditions, we refer to the usual integrability and weak differentiability conditions that are required for (1.4) to hold (see Appendix 7.9 for details).

2. Computing the divergence of matrix valued spectral functions

We present herein a framework to compute the *df* for matrix estimators of the form $S(Y; f)$. Towards this end, we will need to compute the divergence $\nabla \cdot S(Y; f)$, by making use of results due to [28]. For a symmetric matrix $X \in \mathbb{R}^{N\times N}$, let $\lambda_1(X) > \cdots > \lambda_q(X)$ be the set of its unique eigenvalues, $r_1, \ldots, r_q$ be the associated multiplicities, and $E_1(X) \in \mathbb{R}^{N\times r_1}, \ldots, E_q(X) \in \mathbb{R}^{N\times r_q}$ be the set of matrices whose columns are the corresponding orthonormal eigenvectors. For any given function $f : \mathbb{R} \to \mathbb{R}$, define the associated matrix valued function $F : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ as follows:

$$F(X) = \sum_{k=1}^{q} f(\lambda_k(X))E_k(X)E_k(X)'. \tag{2.1}$$

Shapiro [28] investigates differentiability properties of the function $F(X)$ in cases where $f(x)$ is directionally differentiable. Shapiro's study seems to be motivated by the works of [31, 26] on the semismoothness of $F(X)$ when $f(x) = |x|$ or $\max\{0, x\}$, which play important roles in algorithms for semidefinite programs and complementarity problems. For our purpose, we consider a special case of the directional differentiability property of $F(X)$ from [28].

Suppose f is directionally differentiable at every point $\lambda_k(X), k = 1, \ldots, q$. Then the directional derivative $f'(\lambda_k(X); h)$ exists $\forall h \in \mathbb{R}$. Let $\Psi_k : \mathbb{R}^{r_k\times r_k} \to \mathbb{R}^{r_k\times r_k}$ be the associated matrix valued function defined through $f'(\lambda_k(X); \cdot)$. That is, for a given symmetric matrix $Y \in \mathbb{R}^{r_k\times r_k}$,

$$\Psi_k(Y) = \sum_i f'(\lambda_k(X); \lambda_i(Y))E_i(Y)E_i(Y)',$$

where $\{\lambda_i(Y)\}, \{E_i(Y)\}$ are the sets of unique eigenvalues and the corresponding orthonormal eigenvectors of Y, respectively.

Lemma 1. *(Shapiro [28]) Using the notation above, $F(X)$ is directionally differentiable at X and its directional derivative $F'(X;H)$ is given by:*

$$\begin{aligned} F'(X;H) &= \lim_{t \downarrow 0} \frac{F(X+tH)-F(X)}{t} \\ &= \frac{1}{2} \sum_{\substack{l \neq k \\ l,k=1}}^{q} \frac{f(\lambda_l(X)) - f(\lambda_k(X))}{\lambda_l(X) - \lambda_k(X)} (E_l E_l' H E_k E_k' + E_k E_k' H E_l E_l') \\ &\quad + \sum_{k=1}^{q} E_k[\Psi_k(E_k' H E_k)] E_k' \,, \end{aligned} \tag{2.2}$$

where $H \in \mathbb{R}^{N \times N}$ is an arbitrary real symmetric matrix, X is symmetric and E_k denotes $E_k(X)$ for $k = 1, 2, \ldots, q$.

Lemma 1 ensures that matrix valued functions inherit directional differentiability (at a matrix point X which is symmetric), from the real valued function $f(\cdot)$ (at all the distinct eigenvalues of X). Lemma 2 presents a generalization of Lemma 1 to asymmetric matrices—this will be useful to address the differentiability properties of (rectangular) matrix valued spectral functions (see the definition in Section 1.1).

Lemma 2. *For any matrix $Y \in \mathbb{R}^{m \times n}$, consider the reduced singular value decomposition $Y = U \Sigma V'$ with $\Sigma \in \mathbb{R}^{n \times n}$. Thus, there exists $\bar{U} \in \mathbb{R}^{m \times (m-n)}$ such that $\bar{U}'\bar{U} = I \in \mathbb{R}^{(m-n)\times(m-n)}$ and $\bar{U}'U = 0 \in \mathbb{R}^{(m-n)\times n}$. Define the matrices*

$$Y^* = \begin{bmatrix} 0 & Y \\ Y' & 0 \end{bmatrix}, \quad P = \begin{bmatrix} \frac{1}{\sqrt{2}}U & \frac{1}{\sqrt{2}}U & \bar{U} \\ \frac{1}{\sqrt{2}}V & \frac{-1}{\sqrt{2}}V & 0 \end{bmatrix} \quad \text{and} \quad \Sigma^* = \begin{bmatrix} \Sigma & 0 & 0 \\ 0 & -\Sigma & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

An eigendecomposition of Y^* is given by: $Y^* = P \Sigma^* P'$.

The relation between the singular value decomposition of a matrix Y and the Schur decomposition of its symmetrized version Y^* is a well known result in matrix-theory – see [14] for example. In our case, Lemma 2 provides a tool to study the directional differentiability of matrix valued spectral functions via Lemma 1. In particular, for any given $S(Y;f)$, we can define a real valued function $f^* : \mathbb{R} \to \mathbb{R}$ as $f^*(x) = f(x)$ for $x \geq 0$ and $f^*(x) = -f(-x)$ otherwise. Let Y^* be the matrix defined in Lemma 2 and $F^*(Y^*)$ be the matrix valued function associated with $f^*(x)$ as described in (2.1). Then Lemma 2 leads to

$$F^*(Y^*) = \begin{bmatrix} 0 & S(Y;f) \\ S(Y;f)' & 0 \end{bmatrix}.$$

Hence the directional differentiability of $S(Y;f)$ can be analyzed by studying the symmetric matrix valued function $F^*(Y^*)$ through Lemma 1. The divergence

of $S(Y;f)$ can then be accordingly derived. An expression for the divergence of matrix valued spectral functions is given in Corollary 1; and the proof is presented in Appendix 7.1.

Corollary 1. *Given a matrix $Y \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \ldots \geq \sigma_n$, let $s_1 > s_2 > \ldots > s_K \geq 0$ be the set of distinct singular values, $d_1, \ldots, d_K$ be the associated multiplicities. For any $f : \mathbb{R}^+ \to \mathbb{R}$ with $f(0) = 0$, if it is differentiable at every point $s_i, 1 \leq i \leq K$, then*

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial[S(Y;f)]_{ij}}{\partial Y_{ij}} = \sum_{s_i>0}\left[\frac{d_i(d_i+1)}{2}f'(s_i) + \left((m-n)d_i + \frac{d_i(d_i-1)}{2}\right)\frac{f(s_i)}{s_i}\right]$$
$$+d_K(m-n+d_K)f'(0)\mathbb{1}(s_K=0) + \sum_{1\leq i\neq j\leq K} d_i d_j \frac{s_i f(s_i) - s_j f(s_j)}{s_i^2 - s_j^2}.$$

We remark that the differentiability condition on f can be weakened to directional differentiability leading to a more complex divergence formula, as derived in Appendix 7.1. We choose to present the streamlined version in Corollary 1 for simplicity. The divergence expression in Corollary 1 appears in earlier work [4]—in this paper, the authors first derive the divergence formula for a matrix Y which is simple and has full rank. Their derivation is based on standard techniques of computing the Jacobian of the SVD [8, 27]. They then extend the result to general matrices. Here we show that the divergence formula can be derived as a consequence of Lemma 1, and can be generalized to a larger class of functions f.

On a related note, the differentiability properties of singular values of a rectangular matrix have been studied in [18, 19, 6]. These results however, are not applicable to our current setting because we are concerned with matrix functions that involve *both* singular values and singular vectors.

3. Degrees of freedom for additive Gaussian models

We start by considering the canonical additive Gaussian model:

$$Y = M^* + \mathcal{E}, \tag{3.1}$$

where $Y \in \mathbb{R}^{m\times n}$ is the observed matrix, $M^* \in \mathbb{R}^{m\times n}$ is the underlying low rank matrix of interest, and $\mathcal{E} = (\epsilon_{ij})_{m\times n}$ is the random noise matrix with $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0,\tau^2)$.

3.1. Estimators obtained via spectral regularization

A popular class of low rank matrix estimators are obtained via spectral regularization:

$$S_\theta(Y) \in \operatorname*{arg\,min}_{M\in\mathbb{R}^{m\times n}} \frac{1}{2}\|Y-M\|_F^2 + \sum_{i=1}^{n} P_\theta(\sigma_i), \tag{3.2}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ are the singular values of M and $P_\theta : \mathbb{R}^+ \to \mathbb{R}^+$ is a family of sparsity promoting penalty functions indexed by θ. For example, $P_\theta(x) = \theta x$ gives the *nuclear norm* penalty [3]. Some non-convex penalty functions include MC+ [40] and SCAD [10]. The optimization problem in (3.2) is closely related to the following problem

$$s_\theta(\boldsymbol{\sigma}) \in \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \frac{1}{2}\|\boldsymbol{\sigma}(Y) - \boldsymbol{\alpha}\|_2^2 + \sum_{i=1}^n P_\theta(\alpha_i), \tag{3.3}$$

where $s_\theta(\boldsymbol{\sigma}) = (s_\theta(\sigma_1), \ldots, s_\theta(\sigma_n)) \in \mathbb{R}^n, \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$, and $\boldsymbol{\sigma}(Y) = (\sigma_1(Y), \ldots, \sigma_n(Y)) \in \mathbb{R}^n$ are the singular values of Y. Due to the separability in (3.3), it is clear that $s_\theta(\cdot)$ is the proximal function induced by the penalty P_θ:

$$s_\theta(u) \in \arg\min_{x \in \mathbb{R}} \ \frac{1}{2}(x-u)^2 + P_\theta(x).$$

Problem (3.2) in fact, admits a closed form solution (See Proposition 1 in [23]):

$$S_\theta(Y) = U\mathrm{diag}(s_\theta(\boldsymbol{\sigma}))V',$$

where $Y = U\mathrm{diag}(\boldsymbol{\sigma})V'$ is the reduced SVD of Y. Since the penalty function $P_\theta(\cdot)$ shrinks some singular values to zero, it results in a low rank matrix estimator $S_\theta(Y)$. An appropriate amount of shrinkage (that is, θ) can be obtained by using the SURE framework—to this end, the following corollary presents SURE expressions for a variety of estimators.

Corollary 2. *Consider the spectral regularized estimator $S_\theta(Y)$ in (3.2) under the model (3.1). Assuming $P_\theta(\cdot)$ is differentiable on $(0, \infty)$ and $P_\theta(0) = 0$, we introduce the following quantity (ϕ_P) that measures the amount of concavity of $P_\theta(\cdot)$:*

$$\phi_P := \inf_{\alpha, \alpha' > 0} \ \frac{P'_\theta(\alpha) - P'_\theta(\alpha')}{\alpha - \alpha'},$$

where $P'_\theta(\alpha)$ denotes the derivative of $P_\theta(\alpha)$ wrt α on $\alpha > 0$. Suppose $\phi_P + 1 > 0$, then

$$df(S_\theta(Y)) = \mathbb{E}\left[\sum_{i=1}^n \left(s'_\theta(\sigma_i) + (m-n)\frac{s_\theta(\sigma_i)}{\sigma_i}\right) + 2\sum_{\substack{i \neq j \\ i,j=1}}^n \frac{\sigma_i s_\theta(\sigma_i)}{\sigma_i^2 - \sigma_j^2}\right], \tag{3.4}$$

where $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ are the singular values of Y.

In a recent piece of work [15], Hansen derives the same *df* formula as in (3.4). However, the result in [15] holds for a different class of matrix estimators than the one considered in Corollary 2. Specifically, Theorem 1 in [15] requires the function $s_\theta(\cdot)$ to be differentiable but allows different $s_\theta(\cdot)$ applied to each of the singular value. In contrast, Corollary 2 assumes the same $s_\theta(\cdot)$ across the singular values, but we allow for a non-differentiable function $s_\theta(\cdot)$. We discuss a few examples below.

The condition $\phi_P + 1 > 0$ holds for many penalty functions. First of all, any convex function differentiable over $(0, \infty)$, has a non-negative ϕ_P. In particular, for $P_\theta(\alpha) = \theta\alpha$, it is straightforward to confirm that $s_\theta(\sigma) = (\sigma - \theta)_+$ (i.e., the soft-thresholding operator). This recovers the *df* formula of the singular value thresholding estimator studied in [4]. Moreover, some families of non-convex penalty functions satisfy $\phi_P + 1 > 0$ as well. Examples include MC+ (for $\gamma > 1$) and SCAD (for $a > 2$), where γ, a are tuning parameters associated with the two penalty functions, respectively. We refer the reader to Section 5 for explicit expressions. Non-convex penalties are well known to attenuate the estimation bias caused by convex sparsity-promoting functions [10, 21, 23]. Note that some popular non-convex penalties like $P_\theta(\alpha) = \theta|\alpha|^q$ $(0 \le q < 1)$ do not satisfy the condition $\phi_P + 1 > 0$. In particular, when $q = 0$, $P_\theta(\alpha) = \theta\mathbb{1}(\alpha \neq 0)$ gives the widely known rank regularized estimator

$$S_\theta(Y) = \sum_{i=1}^{n} \sigma_i \mathbb{1}(\sigma_i > \sqrt{2\theta}) \boldsymbol{u}_i \boldsymbol{v}_i'. \tag{3.5}$$

Due to the hard thresholding rule on the singular values, $S_\theta(Y)$ is not a continuous function of Y, hence Stein's Lemma cannot be directly applied. The following corollary (the proof is presented in Appendix 7.4) derives an expression for the degrees of freedom of the rank regularized estimator.

Corollary 3. *Consider the rank regularized matrix estimator in* (3.5) *under the model* (3.1)*, then*

$$\begin{aligned} df(S_\theta(Y)) = \mathbb{E} \sum_{\substack{i \neq j \\ i,j=1}}^{n} \Big(\frac{\sigma_i^2 \mathbb{1}(\sigma_i > \sqrt{2\theta})}{\sigma_i^2 - \sigma_j^2} + \frac{\sigma_j^2 \mathbb{1}(\sigma_j > \sqrt{2\theta})}{\sigma_j^2 - \sigma_i^2} \Big) \\ + \sum_{i=1}^{n} \Big[(m - n + 1) P(\sigma_i > \sqrt{2\theta}) + \sqrt{2\theta} f_{\sigma_i}(\sqrt{2\theta}) \Big], \end{aligned} \tag{3.6}$$

where $f_{\sigma_i}(\cdot)$ is the marginal probability density function of σ_i, which is the ith singular value of Y.

If we ignore the regularity conditions and use Equation (1.4) directly, we will get an incorrect estimate of the *df* — specifically, the expression we obtain (by applying Corollary 1) will not include the term $\sum_{i=1}^{n} \sqrt{2\theta} f_{\sigma_i}(\sqrt{2\theta})$ appearing in (3.6).

To arrive at (3.6) we construct a sequence of matrix valued spectral functions (induced by MC+ penalty) which satisfy the conditions of Corollary 2. The *df* of this sequence converges to the *df* of the rank regularized matrix estimator. We then combine the formula in Corollary 2 with a careful limiting argument that hinges on the eigenvalue distribution of a central Wishart matrix to derive the *df* of the rank regularized estimator.

Note that when $P_\theta(\alpha) = \theta|\alpha|^q$ with $0 < q < 1$, problem (3.3) does not admit an explicit solution. Introducing the notation

$$\eta_q(\sigma;\theta) = \arg\min_{x\in\mathbb{R}} \frac{1}{2}|\sigma - x|^2 + \theta|x|^q, \tag{3.7}$$

we have $s_\theta(\sigma) = \eta_q(\sigma;\theta)$. According to Lemmas 5 and 6 in [41], the function $\eta_q(\sigma;\theta)$ has a jump discontinuity:

$$\begin{aligned}
&\eta_q(\sigma;\theta) = 0, \text{ for } 0 \le \sigma < c_q\theta^{1/(2-q)},\\
&\eta_q(c_q\theta^{1/(2-q)};\theta) = [2(1-q)\theta]^{1/(2-q)},\\
&c_q = [2(1-q)]^{1/(2-q)} + q[2(1-q)]^{(q-1)/(2-q)}.
\end{aligned} \tag{3.8}$$

Hence $s_\theta(\sigma)$ is not continuous. Similar to the rank regularized estimator, Stein's Lemma is not applicable to the case $P_\theta(\alpha) = \theta|\alpha|^q$ for $q \in (0,1)$. We adapt the approach used in the proof of Corollary 3 to derive the *df* for the case $0 < q < 1$ – the result is presented in the following corollary, the proof of which is in Appendix 7.5.

Corollary 4. *Consider the matrix estimator $S_\theta(Y)$ in (3.2) with $P_\theta(\alpha) = \theta|\alpha|^q$ for $q \in (0,1)$ under the model (3.1), then*

$$\begin{aligned}
df(S_\theta(Y)) = \mathbb{E}\sum_{\substack{i\neq j\\ i,j=1}}^{n} \frac{\sigma_i\eta_q(\sigma_i;\theta) - \sigma_j\eta_q(\sigma_j;\theta)}{\sigma_i^2 - \sigma_j^2}\\
+\mathbb{E}\sum_{i=1}^{n}\left[(m-n)\frac{\eta_q(\sigma_i;\theta)}{\sigma_i} + \eta_q'(\sigma_i;\theta) + [2(1-q)\theta]^{1/(2-q)} f_{\sigma_i}(c_q\theta^{1/(2-q)})\right],
\end{aligned}$$

where $f_{\sigma_i}(\cdot)$ is the marginal density function of Y's ith singular value σ_i; $\eta_q'(\sigma_i;\theta)$ is the partial derivative of $\eta_q(\sigma_i;\theta)$ with respect to σ_i.

By a quick inspection, we observe that setting $q = 1$ and $q = 0$ in the *df* formula of Corollary 4 recovers the *df* formulae for the cases $q = 1$ and $q = 0$— these special cases are already derived in Corollaries 2 and 3 respectively. Hence Corollary 4 presents a unified *df* formula for the family of penalty functions for all $q \in [0,1]$. Note also, that the term $\sum_{i=1}^{n}[2(1-q)\theta]^{1/(2-q)} f_{\sigma_i}(c_q\theta^{1/(2-q)})$ in the expression of Corollary 4 will not appear if we had directly applied Stein's Lemma and Corollary 1 to derive the *df*. Supporing simulation results are presented in Figure 1.

3.2. Reduced rank estimators

We now consider rank constrained estimators [14] of the form:

$$C_K(Y) \in \arg\min_{\text{rank}(M)\le K} \|Y - M\|_F^2, \tag{3.9}$$

for some positive integer $K \le n$. The Eckart-Young Theorem [7] states that

$$C_K(Y) = \sum_{i=1}^{K} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i',$$

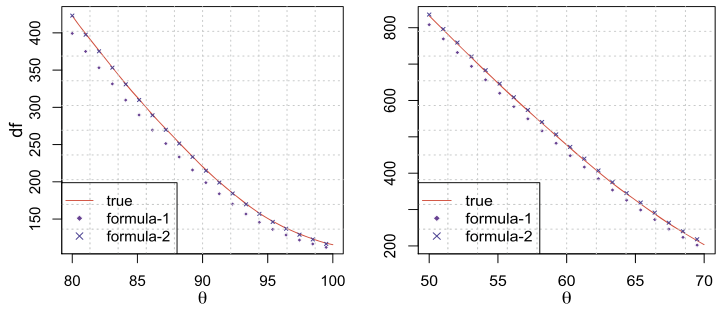

FIG 1. *The df computation of $S_\theta(Y)$ with $P_\theta(\alpha) = \theta|\alpha|^q$ for $q = 0$ (left), and $q = 0.1$ (right). The true df (red curve) is computed according to the definition of df in (1.2). The legend "formula-1" (purple diamond) denotes the df obtained by using (1.4) directly; the legend "formula-2" (blue cross) represents the df derived from Corollary 4. In this simulation, we set $n = m = 50, M^* = 5 \cdot \mathbf{11}'$ (where, $\mathbf{1}$ is a vector of all ones), and $\tau = 1$. The df is calculated by Monte Carlo simulation over 10,000 replications.*

where $\sigma_1 \geq \cdots \geq \sigma_K$ are the K largest singular values of Y, and $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^K$ are the corresponding singular vectors. Here, K controls the amount of regularization. The choice of K can be guided by an expression for the *df* of $C_K(Y)$, as presented below.

Corollary 5. *For the reduced rank estimator $C_K(Y)$ in (3.9) under the model (3.1), we have*

$$df(C_K(Y)) = \begin{cases} \mathbb{E}\Big[(m+n-K)K + 2\sum_{i=1}^{K}\sum_{j=K+1}^{n} \frac{\sigma_j^2}{\sigma_i^2 - \sigma_j^2}\Big], & \text{if } K < n \\ mn & \text{if } K = n \end{cases} \tag{3.10}$$

where $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ are the singular values of Y.

The proof of Corollary 5 can be found in Appendix 7.6. The term $(m+n-K)K$ appearing in the formula of *df* equals the number of free parameters in the specification of a $m \times n$ matrix with rank K. Corollary 5 demonstrates that the degrees of freedom of $C_K(Y)$ is typically larger than the number of free parameters (when $K < n$).

The expression inside the expectation in (3.10) has been proved equal to the divergence $\nabla \cdot C_K(Y)$ in Yuan [38]. This was obtained by fairly involved tools in calculus and tedious algebraic derivations. As we show in Appendix 7.6, we obtain this expression via a direct application of Corollary 2. More importantly, Corollary 5 establishes that $\nabla \cdot C_K(Y)$ is unbiased for $df(C_K(Y))$ – that is, formula (1.4) holds for the matrix estimator $C_K(Y)$. Yuan [38] on the other hand, *assumes* that (1.4) is applicable—here, we present a formal justification of this assumption. The following section provides some intuition regarding the regularity conditions that imply the validity of (1.4).

3.2.1. Verifying the regularity conditions

We showed in Section 3.1 that Stein's Lemma (1.4) does *not* apply to the discontinuous rank regularized estimator $S_\theta(Y)$. In light of this observation, it is important to investigate if $C_K(Y)$ satisfies the regularity conditions sufficient for the identity $df(C_K(Y)) = \mathbb{E}[\nabla \cdot C_K(Y)]$ in (1.4) to hold true. As we discuss below, verifying the weak differentiability of $C_K(Y)$ does not appear to be straightforward.

Firstly, $C_K(Y)$ might not be continuous at Y when $\sigma_K(Y) = \sigma_{K+1}(Y)$. This can be seen by a simple example. Suppose $m = n = 3, K = 2$ and $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ is a set of orthonormal bases in $\mathbb{R}^3$. For $Y = 2\boldsymbol{e}_1\boldsymbol{e}_1' + \boldsymbol{e}_2\boldsymbol{e}_2' + \boldsymbol{e}_3\boldsymbol{e}_3'$, consider a sequence

$$Y_\ell = 2\boldsymbol{e}_1\boldsymbol{e}_1' + (1 + 1/\ell)\boldsymbol{e}_2\boldsymbol{e}_2' + (1 - 1/\ell)\boldsymbol{e}_3\boldsymbol{e}_3' \to Y, \quad \text{as} \quad \ell \to \infty.$$

One can directly verify that as $\ell \to \infty$,

$$C_K(Y_\ell) = 2\boldsymbol{e}_1\boldsymbol{e}_1' + (1 + 1/\ell)\boldsymbol{e}_2\boldsymbol{e}_2' \to 2\boldsymbol{e}_1\boldsymbol{e}_1' + \boldsymbol{e}_2\boldsymbol{e}_2'.$$

Now consider another sequence converging to Y:

$$\tilde{Y}_\ell = 2\boldsymbol{e}_1\boldsymbol{e}_1' + (1 - 1/\ell)\boldsymbol{e}_2\boldsymbol{e}_2' + (1 + 1/\ell)\boldsymbol{e}_3\boldsymbol{e}_3' \to Y, \quad \text{as} \quad \ell \to \infty.$$

For sequence $\tilde{Y}_\ell$ it is clear that as $\ell \to \infty$,

$$C_K(\tilde{Y}_\ell) = 2\boldsymbol{e}_1\boldsymbol{e}_1' + (1 + 1/\ell)\boldsymbol{e}_3\boldsymbol{e}_3' \to 2\boldsymbol{e}_1\boldsymbol{e}_1' + \boldsymbol{e}_3\boldsymbol{e}_3'.$$

Moreover, $C_K(Y)$ might not be Lipschitz continuous over the open ball outside the set $\{Y : \sigma_K(Y) = \sigma_{K+1}(Y)\}$. To illustrate this, we take a simple example as follows. Let $m = n = 2K$ for a positive integer K, and set

$$Y_1 = U\mathrm{diag}(\boldsymbol{\sigma})V', \qquad Y_2 = U\mathrm{diag}(\tilde{\boldsymbol{\sigma}})V',$$

$$\sigma_i = a, \tilde{\sigma}_i = b, \sigma_j = b, \tilde{\sigma}_j = a, 1 \le i \le K, K + 1 \le j \le n,$$

where $U, V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and a, b are two constants satisfying $0 < b < a$. We can then compute that $\|Y_1 - Y_2\|_F^2 = 2K(a-b)^2$, and $\|C_K(Y_1) - C_K(Y_2)\|_F^2 = 2Ka^2$. Hence, by choosing $b = a - 1$, we can conclude

$$\sup_a \frac{\|C_K(Y_1) - C_K(Y_2)\|_F}{\|Y_1 - Y_2\|_F} = \infty.$$

3.2.2. Estimating df via smoothing with convolution operators

The discussions in Section 3.2.1, suggest challenges in legitimately invoking Stein's Lemma to obtain an expression for *df*. We thus pursue a different approach, which to our knowledge, is novel. To this end, we first compute the *df* for a smoothed version of $C_K(Y)$, obtained by the following convolution operation:

$$g_h(Y) = \mathbb{E}_Z[C_K(Y + hZ)],$$

where the elements of $Z \in \mathbb{R}^{m\times n}$ are i.i.d from $N(0,1)$, independent of Y; the expectation $\mathbb{E}_Z(\cdot)$ is taken with respect to Z; and $h > 0$ is a constant. Because $g_h(Y)$ satisfies the regularity conditions, we can derive $df(g_h(Y))$ by computing the divergence of $g_h(Y)$. We show that $df(g_h(Y)) \to df(C_K(Y))$ as $h \to 0+$—using this result, we obtain $df(C_K(Y))$ by letting $h \to 0+$. The detailed analysis is quite technical, and is presented in the Appendix 7.6.

As we were preparing the paper, we became aware of the recent work by Hansen [15] who also provides a rigorous derivation of the *df* for reduced rank estimators. However, the proof technique in [15] is significantly different from ours. Hansen directly verifies the weak differentiability of the estimator and proceeds with divergence calculation—our approach however, uses a continuity argument as explained in the preceding paragraph. Moreover, the approximation strategy via convolution with Gaussian kernel discussed above can work beyond matrix estimation settings; and is hence of independent interest. For example, under the linear regression model, the best subset selection in constrained form is:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 \ \text{ subject to } \|\boldsymbol{\beta}\|_0 \le k.$$

In the orthogonal design setting, the ith coordinate of $\hat{\boldsymbol{\beta}}$ i.e., $\hat{\beta}_i = \boldsymbol{x}_i'\boldsymbol{y}\cdot\mathbb{1}(|\boldsymbol{x}_i'\boldsymbol{y}| \ge |\boldsymbol{x}'\boldsymbol{y}|_{(k)})$, where $\boldsymbol{x}_i$ is the ith column of X and $|\boldsymbol{x}'\boldsymbol{y}|_{(k)}$ is the kth largest value among $\{|\boldsymbol{x}_i'\boldsymbol{y}|\}_i$. The *df* of $\hat{\boldsymbol{\beta}}$ when the underlying signal is null, has been derived in [37] by making use of the projection property of least square estimates. Alternatively, we can follow the continuity argument outlined above, and study the sequence

$$\hat{\beta}_i^h = \mathbb{E}_{\boldsymbol{z}}[(\boldsymbol{x}_i'\boldsymbol{y} + z_i)\cdot\mathbb{1}(|\boldsymbol{x}_i'\boldsymbol{y} + z_i| \ge |\boldsymbol{x}'\boldsymbol{y} + \boldsymbol{z}|_{(k)})], \quad \boldsymbol{z} \sim N(\boldsymbol{0}, h\cdot\mathbf{I}_p)$$

to obtain the *df* formula. Since the calculation is standard, we skip it here.

4. Degrees of freedom in multivariate linear regression

Low rank matrix estimation problems also arise in the multivariate linear regression setting, where one is interested in modeling several response measurements simultaneously. In particular, the multivariate linear regression model is given by:

$$Y = XM^* + \mathcal{E}, \tag{4.1}$$

where, $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)' \in \mathbb{R}^{m\times n}$ is the response matrix, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)' \in \mathbb{R}^{m\times p}$ is the design matrix, $M^* \in \mathbb{R}^{p\times n}$ is the underlying coefficient matrix, and $\mathcal{E} = (\epsilon_{ij})_{m\times n}$ with $\epsilon_{ij} \overset{iid}{\sim} N(0,\tau^2)$ is the random noise matrix.

4.1. Reduced rank regression estimators

In many applications, it is reasonable to assume that the dependency of Y on X is only through $K < \min(p,n)$ linear combinations, namely, M^* is of

low rank. In such cases, we can consider the following reduced rank regression estimator [1, 35]

$$M_K(Y) \in \underset{\text{rank}(M)\leq K}{\arg\min} \|Y - XM\|_F^2. \tag{4.2}$$

Let the compact singular value decomposition of X be $X_{m\times p} = U_{m\times r}\Sigma_{r\times r}V'_{p\times r}$, with r being the rank of X. Then the least squares fit is given by

$$\hat{Y} = X(X'X)^+X'Y = UU'Y, \tag{4.3}$$

where $(X'X)^+$ is the Moore-Penrose pseudo inverse of $X'X$. By applying the Eckart-Young Theorem, an explicit solution of (4.2) is given as follows [38, 25]:

$$M_K(Y) = (X'X)^{-1}X'C_K(\hat{Y}) \quad \text{if } r = p < m.$$

$M_K(Y)$ might not be unique when $p > m$, but the fitted value $XM_K(Y)$ is unique with $XM_K(Y) = C_K(UU'Y)$. The reduced rank problem (3.9) can be thought of as a special case of (4.2) where X equals the identity matrix $\mathbf{I} \in \mathbb{R}^{m\times m}$. We will use the *df* result for the reduced rank estimator in Corollary 5 to derive the *df* formula for the estimator defined in (4.2). It is important to note that, in the current regression setting, we are interested in the prediction error $\mathbb{E}(\|XM^* - XM_K(Y)\|_F^2)$ rather than the estimation error $\mathbb{E}(\|M^* - M_K(Y)\|_F^2)$. Therefore, the degrees of freedom for $M_K(Y)$ is defined as

$$df(M_K(Y)) = \sum_{ij} \text{Cov}((XM_K(Y))_{ij}, Y_{ij})/\tau^2,$$

where $(XM_K(Y))_{ij}$ is the (i,j)th entry of the matrix $XM_K(Y)$.

Corollary 6. *Consider the reduced rank regression estimator $M_K(Y)$ in (4.2) under the model (4.1). We have the following df formula for $M_K(Y)$:*

$$df(M_K(Y)) = \begin{cases} \mathbb{E}\Big[(r+n-K)K + 2\sum_{i=1}^{K}\sum_{j=K+1}^{\min(r,n)} \frac{\sigma_j^2}{\sigma_i^2-\sigma_j^2}\Big], & \text{if } K < \min(r,n) \\ rn, & \text{if } K \geq \min(r,n) \end{cases} \tag{4.4}$$

where $\sigma_1 \geq \ldots \geq \sigma_{\min(r,n)} \geq 0$ are the singular values of the least squares fitted value $\hat{Y}$, as defined in (4.3).

Note that the analytic expression inside the expectation in (4.4) has been shown to be equal to $\nabla \cdot (XM_K(Y))$ in Yuan [38], Mukherjee et al. [25]. Both papers use the chain rule to relate the divergence of $XM_K(Y)$ to the divergence of a related reduced rank estimator. Our approach differs as we compute the *df* from basic principles and then appeal to Corollary 5. We emphasize that it is not immediately clear if the regularity conditions sufficient for the identity in (1.4) to hold, are satisfied — see also the discussion in Section 3.2.1. Based on the result in Corollary 5, we present a formal justification of the result in (4.4).

4.2. Spectral regularized regression estimators

In addition to the constrained estimator in (4.2), we may also consider the penalized problem

$$\underset{M\in\mathbb{R}^{p\times n}}{\arg\min} \quad \frac{1}{2}\|Y - XM\|_F^2 + \sum_{i=1}^{\min(p,n)} P_\theta(\sigma_i). \tag{4.5}$$

However, unlike the spectral regularized problem (3.2), except for few penalty functions like $P_\theta(\sigma) = \theta\mathbb{1}(\sigma \neq 0)$ [2], there is no closed form solution for (4.5). Simple expressions for the degrees of freedom for such fitting procedures seem to be unknown. We note however, that some nice work is available on the *df* of regularized estimators in the linear regression setting—see for e.g. Zou et al. [42], Tibshirani and Taylor [34].

We follow the approach of Mukherjee et al. [25]. Motivated by the solution form of (4.2), we explicitly construct an estimator for M^* given by

$$RM_\theta(Y) = (X'X)^{-1}X'S_\theta(\hat{Y}), \qquad X \cdot RM_\theta(Y) = S_\theta(UU'Y), \tag{4.6}$$

where $\hat{Y}$ is the least square fitted value, U is the left singular vector matrix of X, and $S_\theta(\cdot)$ is defined in (3.2). The following two corollaries provide an expression of the *df* for a variety of such estimators.

Corollary 7. *For the penalized multivariate regression estimator $RM_\theta(Y)$ in (4.6) under the model (4.1), if the same conditions for $P_\theta(\cdot)$ as in Corollary 2 hold, then*

$$df(RM_\theta(Y)) = \mathbb{E}\left[\sum_{i=1}^{\min(r,n)} \left(s'_\theta(\sigma_i) + |r-n|\frac{s_\theta(\sigma_i)}{\sigma_i}\right) + 2\sum_{\substack{i\neq j\\ i,j=1}}^{\min(r,n)} \frac{\sigma_i s_\theta(\sigma_i)}{\sigma_i^2 - \sigma_j^2}\right],$$

where $\sigma_1 \geq \ldots \geq \sigma_{\min(r,n)} \geq 0$ are the singular values of the least square fitted value $\hat{Y}$ in (4.3).

Corollary 8. *Consider the penalized multivariate regression estimator $RM_\theta(Y)$ in (4.6) with $P_\theta(\alpha) = \theta|\alpha|^q$ for $q \in (0,1)$, under the model (4.1). We have*

$$df(RM_\theta(Y)) = \mathbb{E}\sum_{\substack{i\neq j\\ i,j=1}}^{\min(r,n)} \frac{\sigma_i\eta_q(\sigma_i;\theta) - \sigma_j\eta_q(\sigma_j;\theta)}{\sigma_i^2 - \sigma_j^2} \quad +$$

$$\mathbb{E}\sum_{i=1}^{\min(r,n)} \left[|r-n|\frac{\eta_q(\sigma_i;\theta)}{\sigma_i} + \eta'_q(\sigma_i;\theta) + [2(1-q)\theta]^{1/(2-q)} f_{\sigma_i}(c_q\theta^{1/(2-q)})\right],$$

where $f_{\sigma_i}(\cdot)$ is the marginal probability density function of the ith singular value of Y (i.e., σ_i); $\eta'_q(\sigma_i;\theta)$ is the partial derivative of $\eta_q(\sigma_i;\theta)$ with respect to σ_i; $\eta(\cdot;\theta), c_q$ are defined in (3.7) and (3.8) respectively; $\sigma_1 \geq \ldots \geq \sigma_{\min(r,n)} \geq 0$ are the singular values of the least square fitted value $\hat{Y}$, appearing in (4.3).

The results in the above two corollaries for $RM_\theta(Y)$ notably differ from that in [25]. We present a formal justification for the unbiasedness of the divergence for $df(RM_\theta(Y))$ under a wide class of non-convex penalties $P_\theta(\cdot)$ in Corollary 7. Furthermore, Corollary 8 presents the *df* formula for a family of penalty functions for which Stein's Lemma does not apply. We refer the reader to Table 1 for a summary of different penalty functions and whether Stein's Lemma applies or not.

TABLE 1

Examples of commonly used penalty functions. We summarize when Stein's Lemma is applicable for the estimator $S_\theta(Y)$ in (3.2) under model (3.1); and the estimator $RM_\theta(Y)$ in (4.6) under model (4.1).

Penalty name	Penalty function	Stein's Lemma
Lasso [32]	$P_\theta(\sigma) = \theta\sigma$	Yes
SCAD [10]	$P_\theta(\sigma) = \begin{cases} \theta\sigma & 0 \le \sigma \le \theta \\ \frac{-\sigma^2+2a\theta\sigma-\theta^2}{2(a-1)} & \theta < \sigma \le a\theta \\ \frac{(a+1)\theta^2}{2} & \sigma > a\theta \end{cases}$	Yes, when $a > 2$
MC+ [40]	$P_\theta(\sigma) = \begin{cases} \theta(\sigma - \frac{\sigma^2}{2\theta\gamma}) & 0 \le \sigma \le \gamma\theta \\ \frac{\gamma\theta^2}{2} & \sigma > \gamma\theta \end{cases}$	Yes, when $\gamma > 1$
Bridge [11]	$P_\theta(\sigma) = \theta\sigma^q, \quad q \in [0,1)$	No
Log [21]	$P_\theta(\sigma) = \frac{\theta \log(\gamma\sigma+1)}{\log(1+\gamma)}$	Yes, when $\frac{\log(1+\gamma)}{\theta\gamma^2} > 1$
Firm [13]	$P_\theta(\sigma) = \begin{cases} \theta(\sigma - \frac{\sigma^2}{2\gamma}) & 0 \le \sigma \le \gamma \\ \frac{\gamma\theta}{2} & \sigma > \gamma \end{cases}$	Yes, when $\gamma > \theta$

5. Simulations

In this section, we perform simulation studies to lend further support to the *df* formulas presented in Sections 3 and 4.

5.1. Additive Gaussian model

We generate synthetic data Y according to the canonical additive Gaussian model (3.1):

$$Y = M^* + \mathcal{E},$$

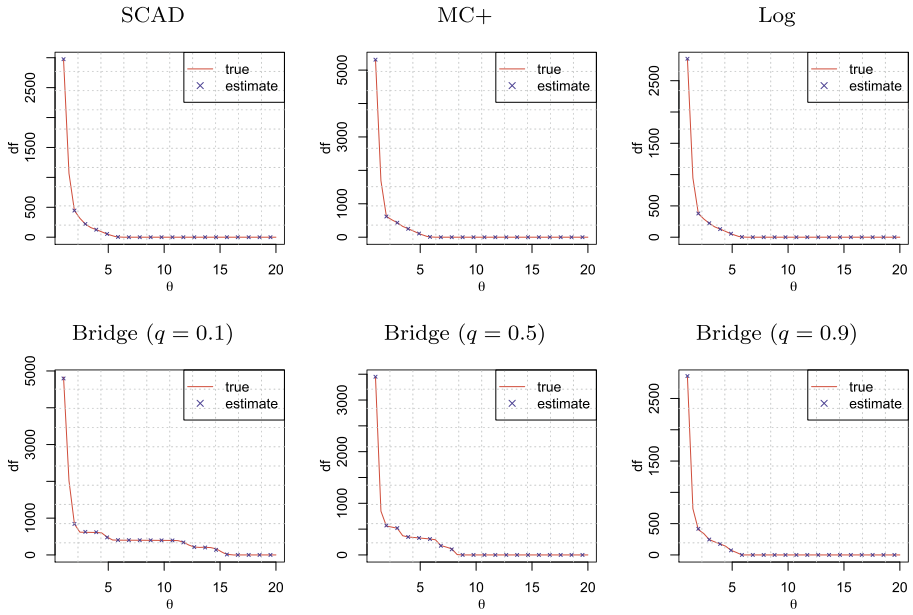

FIG 2. *Degrees of freedom under the additive Gaussian model. The true df (red curve) is computed from (1.2). The estimate (blue cross) is the average of the unbiased estimator over 100 repetitions.*

where $Y \in \mathbb{R}^{m\times n}, M^* \in \mathbb{R}^{m\times n}$, and $\mathcal{E} = (\epsilon_{ij})_{m\times n}$ with $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0,\tau^2)$. We set $m = n = 100, \tau = 0.1, M^* = \sum_{k=1}^{5} k\boldsymbol{u}_k\boldsymbol{u}_k'$, where all entries of the $\boldsymbol{u}_k$'s are independently sampled from $N(0, 1/\sqrt{n})$. We consider the spectral regularized estimator $S_\theta(Y)$ in (3.2) with the following non-convex penalty functions:

(1) The SCAD penalty [10]

$$P_\theta(\sigma) = \theta\sigma\mathbb{1}(\sigma \le \theta) + \frac{-\sigma^2 + 2a\theta\sigma - \theta^2}{2(a-1)}\mathbb{1}(\theta < \sigma \le a\theta) + \frac{(a+1)\theta^2}{2}\mathbb{1}(\sigma > a\theta),$$

where $a > 2$ is a fixed parameter. We choose $a = 3.7$ as used in Fan and Li [10].

(2) The MC+ penalty [40]

$$P_\theta(\sigma) = \theta\Big(\sigma - \frac{\sigma^2}{2\theta\gamma}\Big)\mathbb{1}(0 \le \sigma \le \gamma\theta) + \frac{\gamma\theta^2}{2}\mathbb{1}(\sigma > \gamma\theta),$$

where $\gamma > 0$ is a fixed constant. We set $\gamma = 2$.

(3) The log-penalty [21]

$$P_\theta(\sigma) = \frac{\theta}{\log(1+\gamma)}\log(1+\gamma\sigma), \quad \gamma > 0.$$

We choose $\gamma = 0.01$.

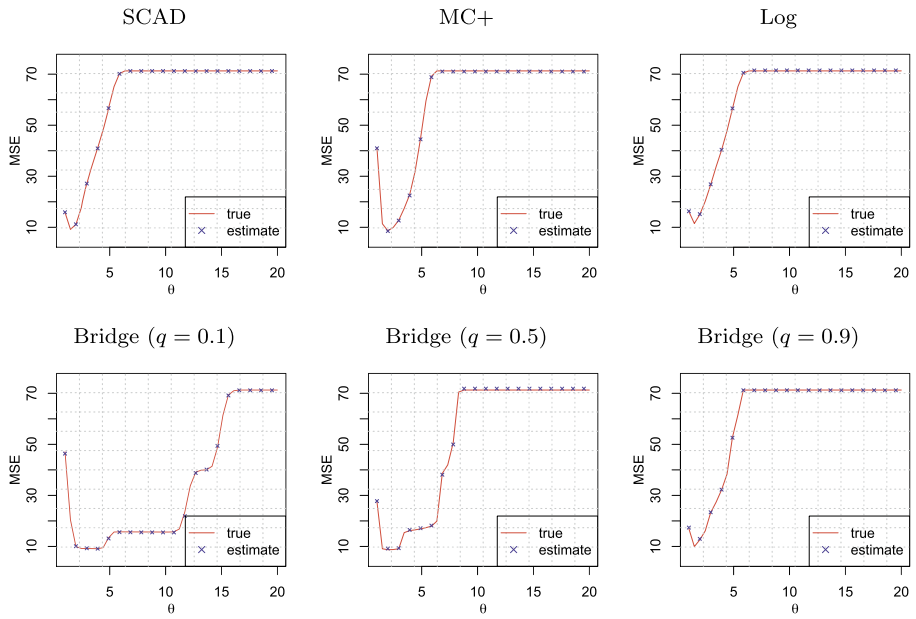

FIG 3. *Expected MSE under the additive Gaussian model. The truth (red curve) is computed via Monte Carlo simulation. The estimate (blue cross) is the average of the unbiased estimator over 100 repetitions.*

(4) The bridge-penalty [11]

$$P_\theta(\sigma) = \theta\sigma^q, \quad q \in [0, 1).$$

We consider $q = 0.1, 0.5, 0.9$.

It is straightforward to verify that the first three penalty functions above satisfy the conditions in Corollary 2 for $\theta \in (0, 20]$. Hence we can use formula (3.4) in Corollary 2, to construct an unbiased estimator for $df(S_\theta(Y))$ when $\theta \in (0, 20]$. For the bridge-penalty function, we use the result in Corollary 4 to obtain the estimator for the *df*. Moreover, for each matrix estimator $S_\theta(Y)$, we compute its *df* (the ground truth) according to the definition (1.2).

Figure 2 depicts the true *df* and its unbiased estimate for the aforementioned non-convex penalties with θ varying over $[0, 20]$. It is clear that the ground truth and the (averaged) estimates are compatible for all the penalties and values of θ under consideration, thus offering empirical support for the correctness of the derived *df* expressions.

In addition to *df*, we further evaluate the estimation of the expected MSE i.e., $\mathbb{E}\|S_\theta(Y) - M^*\|_2^2$. Recall that for a given $S_\theta(Y)$, once an unbiased estimator of the *df* is available, an unbiased estimate for the expected MSE can be constructed based on (1.3). In the present case, we will use the *df* estimates to obtain the estimates for $\mathbb{E}\|S_\theta(Y) - M^*\|_2^2$ according to (1.3). Figure 3 shows the

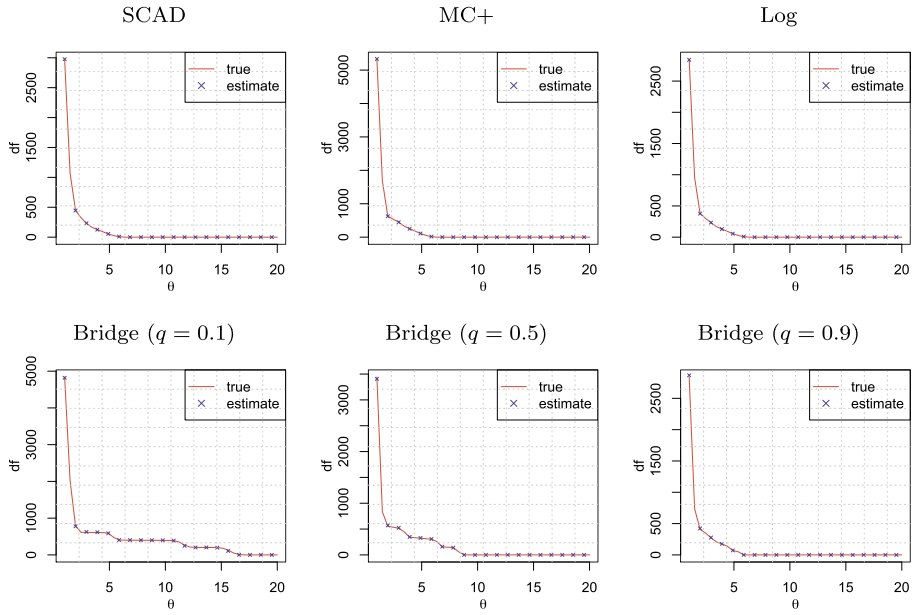

FIG 4. *Degrees of freedom under multivariate linear regression. The true df (red curve) is computed from (1.2). The estimate (blue cross) is the average of the (theoretical) unbiased estimator across100 replications.*

expected MSE and its estimates for the four types of non-convex penalties with $\theta \in [0, 20]$. We observe that the (averaged) estimates are in agreement with the truth.

5.2. *Multivariate linear regression*

We generate data Y according to the multivariate linear regression model (4.1):

$$Y = XM^* + \mathcal{E},$$

where $Y \in \mathbb{R}^{m\times n}, X \in \mathbb{R}^{m\times p}, M^* \in \mathbb{R}^{p\times n}$, and $\mathcal{E} = (\epsilon_{ij})_{m\times n}$ with $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \tau^2)$. We set $m = 300, n = p = 100, \tau = 0.1$, and M^* as in Section 5.1. Each row of the design matrix X is independently sampled from $N(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a Toeplitz matrix with the (i, j)th entry equal to $0.5^{|i-j|}/m$ for $1 \le i, j \le n$. We consider the regularized estimator $RM_\theta(Y)$ in (4.6) with the same non-convex penalty functions studied in Section 5.1. Here, we are interested in *df* in the context of the prediction error $\mathbb{E}\|XM^* - XRM_\theta(Y)\|_2^2$:

$$df(RM_\theta(Y)) = \sum_{ij} \text{Cov}((XRM_\theta(Y))_{ij}, Y_{ij})/\tau^2.$$

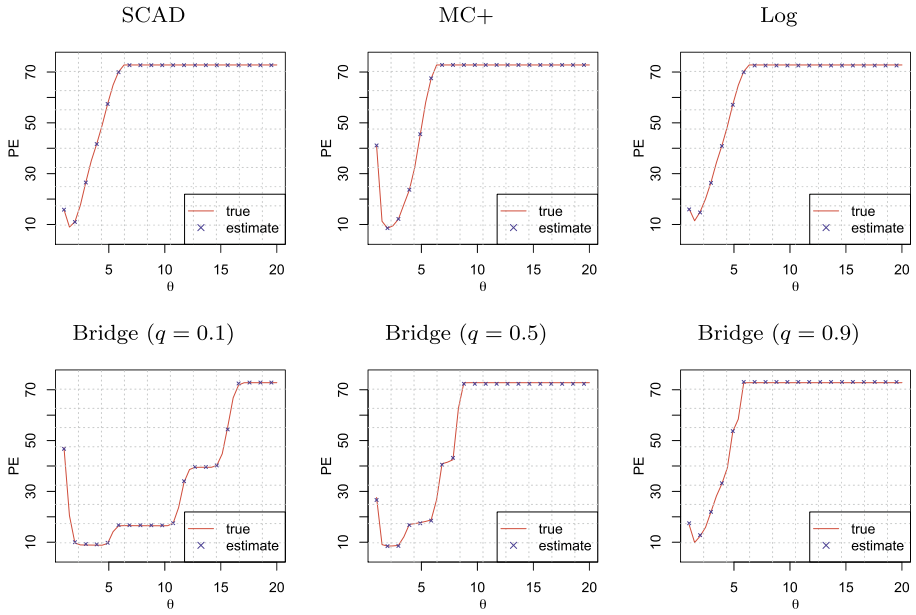

FIG 5. *Prediction error (PE) under multivariate linear regression. The truth (red curve) is computed via Monte Carlo simulation. The estimate (blue cross) is obtained from the (theoretical) unbiased estimator upon averaging over 100 replications.*

Corollaries 7 and 8, lead to estimates of the *df*. As in Section 5.1, we can also construct the estimates for the prediction error (PE) according to (1.3). Figures 4 and 5 depict the comparison between the estimates and the truth for the *df* and PE, respectively. The plots empirically validate the correctness of the *df* expressions derived in Section 4.

6. Conclusion

In this paper, we present a systematic study of computing the degrees of freedom for a wide range of low rank matrix estimators, under the SURE framework. As a building block for the computation, the divergence formula for general spectral functions is derived by appealing to a fundamental result on differentiability of matrix functions due to Shapiro [28]. For a class of estimators, our *df* expressions depend upon the use of Stein's Lemma (i.e., the divergence formula)—in these cases, we have rigorously established the regularity conditions sufficient for Stein's Lemma to be applicable. For other estimators where, Stein's Lemma does not seem to be readily applicable (as the sufficient regularity conditions are not satisfied or difficult to verify), we propose a new Gaussian convolution method and successfully derive their *df* expressions. The estimators studied in this paper include those studied in the recent literature as special cases—our

approach either presents a simple derivation of the *df* expressions and/or complements existing analyses via a rigorous justification for the applicability of Stein's Lemma.

7. Appendix

Here we present proofs of all the main results presented in the paper. The organization is listed below:

1. Appendix 7.1 proves Corollary 1.
2. Appendix 7.2 proves a lemma that is useful in multiple places.
3. Appendix 7.3 proves Corollary 2.
4. Appendix 7.4 proves Corollary 3.
5. Appendix 7.5 proves Corollary 4.
6. Appendix 7.6 proves Corollary 5.
7. Appendix 7.7 proves Corollary 6.
8. Appendix 7.8 proves Corollaries 7 and 8.
9. Appendix 7.9 reviews the regularity conditions that are sufficient for the SURE formula to be applicable.

7.1. *Proof of Corollary 1*

We present a more general result than what appears in Corollary 1, and prove the general result by making use of Lemmas 1 and 2. The proof of Corollary 1 follows as a special case.

Theorem. *Given a matrix $Y \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \ldots \geq \sigma_n$; let $s_1 > s_2 > \ldots > s_K \geq 0$ be the set of distinct singular values, and $d_1, \ldots, d_K$ be the associated multiplicities. Consider a function $f : \mathbb{R}^+ \to \mathbb{R}$ with $f(0) = 0$ that is differentiable at every point s_i with $d_i > 1$ and directionally differentiable at every point s_i with $d_i = 1$. Let $\mathcal{D}$ denote the set of points where f is directionally differentiable but not differentiable. Then*

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial[S(Y;f)]_{ij}}{\partial Y_{ij}} \\
= \quad &\sum_{s_i>0}\left[\frac{d_i(d_i+1)}{2}f'(s_i)\mathbb{1}(s_i\notin\mathcal{D}) + \left((m-n)d_i + \frac{d_i(d_i-1)}{2}\right)\frac{f(s_i)}{s_i}\right] + \\
&d_K(m-n+d_K)f'(0)\mathbb{1}(s_K=0) + \sum_{1\leq i\neq j\leq K} d_i d_j \frac{s_i f(s_i) - s_j f(s_j)}{s_i^2 - s_j^2} + \\
&\sum_{\substack{s_k>0\\ s_k\in\mathcal{D}}}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[u_{ik}^2 v_{jk}^2 f'(s_k;1)\mathbb{1}(u_{ik}v_{jk}>0) - u_{ik}^2 v_{jk}^2 f'(s_k;-1)\mathbb{1}(u_{ik}v_{jk}<0)\right],
\end{aligned}$$

where u_{ik} is the (i,k)th entry of the left singular vector matrix U of Y. A similar notation applies to v_{ik} for the right singular vector matrix V.

According to the above theorem, if f is differentiable at some singular value s_j, the corresponding singular vectors do not appear in the divergence formula of $S(Y;f)$. Under the conditions of Corollary 1, $\mathcal{D}=\emptyset$. This directly leads to the formula appearing in Corollary 1. From the proof that appears below, we can show a more general result: the directional differentiability of f at singular values of Y is sufficient to guarantee the existence of $\nabla\cdot S(Y;f)$. But since the explicit formula is complicated, we skip it for simplicity.

Proof. We focus on the more complicated setting when $s_K=0$. The case in which Y is of full rank can be analyzed in the same way. We first assume f is differentiable at every point $s_j, 1\le j\le K$. Consider the symmetric matrix in Lemma 2: it follows that Y^* has distinct eigenvalues $\pm s_1,\dots,\pm s_{K-1},0$ with multiplicities $d_1,\dots,d_{K-1},2d_K+m-n$. To simplify the notations in the later calculations, we denote those distinct eigenvalues by $\{\mu_k\}_{k=1}^{2K-1}$ and the corresponding multiplicities by $\{r_k\}_{k=1}^{2K-1}$. Define a real function $f^*:\mathbb{R}\to\mathbb{R}$ as $f^*(x)=f(x)$ for $x\ge 0$ and $f^*(x)=-f(-x)$ for $x<0$. Let $F^*(Y^*)$ be the corresponding matrix valued function stated in Lemma 1. The eigenvalue decomposition in Lemma 2 implies a key connection between $F^*(Y^*)$ and $S(Y;f)$

$$F^*(Y^*)=\begin{bmatrix}0 & S(Y;f)\\ S(Y;f)' & 0\end{bmatrix}. \tag{7.1}$$

Let $e_{ij}\in\mathbb{R}^{m\times n}$ be the canonical basis matrix in Euclidean space, i.e., the matrix with all entries equal to 0 but the (i,j)th equal to 1, and denote

$$h_{ij}=\begin{bmatrix}0 & e_{ij}\\ e_{ij}' & 0\end{bmatrix}.$$

Note that (7.1) leads to

$$\lim_{t\downarrow 0}\frac{F^*(Y^*+th_{ij})-F^*(Y^*)}{t}=\begin{bmatrix}0 & \frac{\partial S(Y;f)}{\partial Y_{ij}}\\ \left(\frac{\partial S(Y;f)}{\partial Y_{ij}}\right)^T & 0\end{bmatrix}.$$

By the differentiability of f at $s_j, 1\le j\le K$; f^* is differentiable at all the distinct eigenvalues of Y^*. We can thus apply Lemma 1 to $F^*(Y^*)$ with $H=h_{ij}$. After some algebraic manipulations, we obtain[4]

$$\begin{aligned}&\sum_{i,j}\frac{\partial[S(Y;f)]_{ij}}{\partial Y_{ij}}=\sum_{i,j}\operatorname{tr}\left\{\frac{S(Y;f)}{\partial Y_{ij}}e_{ij}'\right\}\\ &=\frac{1}{2}\sum_{i,j}\sum_{l\ne k,l,k=1}^{q}g_{lk}\Big\{2\operatorname{tr}[E_l(1)E_l(2)'e_{ij}'E_k(1)E_k(2)'e_{ij}']+\\ &\operatorname{tr}[E_l(1)E_l(1)'e_{ij}E_k(2)E_k(2)'e_{ij}']+\operatorname{tr}[E_k(1)E_k(1)'e_{ij}E_l(2)E_l(2)'e_{ij}']\Big\}+\end{aligned}$$

[4]Note that the second term on the right hand side of Equation (2.2) in Lemma 1 is $\sum_{k=1}^{q}f'(\mu_k(X))E_kE_k'HE_kE_k'$

$$\sum_{i,j}\sum_{k=1}^{q}(f^*(\mu_k))'\Big\{\mathrm{tr}[E_k(1)E_k(2)'e'_{ij}E_k(1)E_k(2)'e'_{ij}] \; + \; \mathrm{tr}[E_k(1)E_k(1)'e_{ij}E_k(2)E_k(2)'e'_{ij}]\Big\} \triangleq I + II, \tag{7.2}$$

where $g_{lk} = \frac{f^*(\mu_l)-f^*(\mu_k)}{\mu_l-\mu_k}$, $q = 2K-1$ is the number of distinct eigenvalues and $E_k(1), E_k(2)$ are the first m rows and last n rows of the eigenvector matrix E_k, respectively. We have used I, II to represent the two summations $\frac{1}{2}\sum_{i,j}\sum_{l\neq k,l,k=1}^{q}(\cdot)$ and $\sum_{i,j}\sum_{k=1}^{q}(\cdot)$, respectively. Let $E_k(1) = (\boldsymbol{w}_1^k, \ldots, \boldsymbol{w}_{r_k}^k)$, $E_k(2) = (\boldsymbol{z}_1^k, \ldots, \boldsymbol{z}_{r_k}^k)$ and $w_1^k(i)$ be the ith element of $\boldsymbol{w}_1^k$. We then have

$$\begin{aligned} T(\mu_l,\mu_k) &\triangleq \sum_{ij}\mathrm{tr}[E_l(1)E_l(2)'e'_{ij}E_k(1)E_k(2)'e'_{ij}] \\ &= \sum_{ij}\sum_{a=1}^{r_l}\sum_{b=1}^{r_k}\mathrm{tr}[\boldsymbol{w}_a^l(\boldsymbol{z}_a^l)'e'_{ij}\boldsymbol{w}_b^k(\boldsymbol{z}_b^k)'e'_{ij}] \\ &= \sum_{a=1}^{r_l}\sum_{b=1}^{r_k}\sum_{ij} z_a^l(j)w_b^k(i)z_b^k(j)w_a^l(i) = \sum_{a=1}^{r_l}\sum_{b=1}^{r_k}[(\boldsymbol{w}_b^k)'\boldsymbol{w}_a^l]\cdot[(\boldsymbol{z}_b^k)'\boldsymbol{z}_a^l] \\ &\overset{(a)}{=} \begin{cases} 0 & \text{if } |\mu_k| \neq |\mu_l| \text{ or } |\mu_k\mu_l| = 0 \\ \mathrm{sign}(\frac{\mu_k}{\mu_l})\frac{r_k}{4} & \text{otherwise.} \end{cases} \end{aligned} \tag{7.3}$$

Here, (a) follows due to the fact that $\boldsymbol{w}_a^k$ (and $\boldsymbol{z}_a^k$) is one of the columns of the matrix $(\frac{1}{\sqrt{2}}U, \frac{1}{\sqrt{2}}U, \bar{U})$ (and $(\frac{1}{\sqrt{2}}V, \frac{-1}{\sqrt{2}}V, 0)$, respectively), by checking the eigenvector matrix specified in Lemma 2. Similarly, we also get

$$\begin{aligned} G(\mu_l,\mu_k) &\triangleq \sum_{ij}\mathrm{tr}[E_l(1)E_l(1)'e_{ij}E_k(2)E_k(2)'e'_{ij}] = \sum_{a=1}^{r_l}\sum_{b=1}^{r_k}||\boldsymbol{w}_a^l||^2\cdot||\boldsymbol{z}_b^k||^2 \\ &= \begin{cases} \frac{r_k r_l}{4} & \text{if } \mu_l,\mu_k \neq 0 \\ \frac{r_k(m-n+d_K)}{2} & \text{if } \mu_l = 0, \mu_k \neq 0 \\ \frac{r_l d_K}{2} & \text{if } \mu_l \neq 0, \mu_k = 0 \\ d_K(m-n+d_K) & \text{if } \mu_l = \mu_k = 0. \end{cases} \end{aligned} \tag{7.4}$$

We now use the results (7.3) and (7.4) to calculate I and II in (7.2). Recall that $\{\mu_k\}_{k=1}^q = \{\pm s_1, \ldots, \pm s_{K-1}, 0\}$, $\{r_k\}_{k=1}^q = \{d_1, \ldots, d_K\}$ and f^* is an odd function. It is then not hard to see that

$$\sum_{l\neq k,l,k=1}^{q} g_{lk}T(\mu_l,\mu_k) = \frac{-1}{2}\sum_{s_k>0} d_k\frac{f(s_k)}{s_k}$$

and

$$\frac{1}{2}\sum_{l\neq k,l,k=1}^{q} g_{lk}(G(\mu_l,\mu_k) + G(\mu_k,\mu_l))$$

$$= \sum_{\substack{\mu_l\neq 0,\mu_k\neq 0\\ l\neq k}} g_{lk}\frac{r_k r_l}{4} + \sum_{\mu_l=0,\mu_k\neq 0} g_{lk}\frac{r_k(m-n+2d_K)}{2}$$

$$= \sum_{s_k>0}\frac{f(s_k)}{s_k}\left(\frac{d_k^2}{2}+d_k(m-n)\right) + \sum_{l\neq k,l,k=1}^{K} d_l d_k \frac{f(s_l)s_l - f(s_k)s_k}{s_l^2 - s_k^2}.$$

Therefore,

$$I = \frac{1}{2}\sum_{l\neq k,l,k=1}^{q} g_{lk}(2T(\mu_l,\mu_k)+G(\mu_l,\mu_k)+G(\mu_k,\mu_l)) =$$

$$\sum_{s_k>0}\frac{f(s_k)}{s_k}\left(\frac{d_k(d_k-1)}{2}+d_k(m-n)\right) + \sum_{l\neq k,l,k=1}^{K} d_l d_k \frac{f(s_l)s_l - f(s_k)s_k}{s_l^2 - s_k^2}. \quad (7.5)$$

Regarding II, it is straightforward to do the computation and obtain,

$$\begin{aligned} II &= \sum_{k=1}^{q}(f^*(\mu_k))'(T(\mu_k,\mu_k)+G(\mu_k,\mu_k)) \\ &= \sum_{s_k>0}\frac{d_k}{2}f'(s_k) + d_K(m-n+d_K)f'(0) + \sum_{s_k>0}\frac{d_k^2}{2}f'(s_k). \end{aligned} \quad (7.6)$$

Combining (7.2), (7.5) and (7.6) gives the divergence formula.

When f is only directionally differentiable at some singular value points, the first part I remains the same since it does not involve the directional derivatives. For the second part II, the derivative should be replaced by the directional derivative if the former does not exist. According to the conditions in the theorem, it is sufficient to consider non-zero singular values with multiplicity one at which f is only directionally differentiable. Suppose $s_k > 0$ is one such point. Then f^* will be only directionally differentiable at $\pm s_k$ (note that f^* is always differentiable at 0). Recall that we have used $\{\mu_k\}_{k=1}^q$ to denote all the distinct eigenvalues of $Y^* : \pm s_1, \ldots, \pm s_{K-1}, 0$. If $\mu_a = s_k, \mu_b = -s_k$, then the terms involving μ_a, μ_b in II can be simplified as follows:

$$\begin{aligned} &2\sum_{ij}[E_a'(2)e_{ij}'E_a(1)]^2(f^*)'(s_k;\operatorname{sign}(E_a'(2)e_{ij}'E_a(1)))\cdot\operatorname{sign}(E_a'(2)e_{ij}'E_a(1)) + \\ &2\sum_{ij}[E_b'(2)e_{ij}'E_b(1)]^2(f^*)'(-s_k;\operatorname{sign}(E_b'(2)e_{ij}'E_b(1)))\cdot\operatorname{sign}(E_b'(2)e_{ij}'E_b(1)) \\ =&\sum_{ij}u_{ik}^2 v_{jk}^2 f'(s_k;\operatorname{sign}(u_{ik}v_{jk}))\cdot\operatorname{sign}(u_{ik}v_{jk}), \end{aligned}$$

where we have used the facts that

$$(f^*)'(s_k;h) = -(f^*)'(-s_k;-h), \quad (f^*)'(s_k;h) = f'(s_k;h), \ \forall h\in\mathbb{R},$$
$$E_a(1) = E_b(1) = \frac{\boldsymbol{u}_k}{\sqrt{2}}, \quad \text{and} \quad E_a(2) = -E_b(2) = \frac{\boldsymbol{v}_k}{\sqrt{2}},$$

and the notation $f'(x;h)$ for the directional derivative of a function $f(\cdot)$ at x along the direction h as defined in (1.5). This completes the proof. □

7.2. A useful lemma

We present a lemma below that will be used multiple times in subsequent proofs.

Lemma 3. *Under the canonical additive Gaussian model $Y = M^* + \mathcal{E}$, let the singular values of $Y \in \mathbb{R}^{m\times n}$ be $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$, then we have (1) $\mathbb{E}\left(\frac{\sigma_i}{\sigma_i-\sigma_j}\right) < \infty$, (2) $\mathbb{E}\left(\frac{1}{\sigma_i-\sigma_j}\right) < \infty$, and (3) $\mathbb{E}\left(\frac{\sigma_i^2}{\sigma_i^2-\sigma_j^2}\right) < \infty$, where $1 \leq i < j \leq n$.*

Proof. Firstly, we show the results hold when $M^* = \mathbf{0}$. Let $\lambda_i = \sigma_i^2$ for $1 \leq i \leq n$, then $\lambda_1 \geq \ldots \geq \lambda_n$ are the eigenvalues of $Y'Y$. The joint distribution $f(\lambda_1, \ldots, \lambda_n)$ of the eigenvalues of a real-valued central Wishart matrix, is known to be [24]:

$$f(\lambda_1, \ldots, \lambda_n) \propto \prod_{a=1}^{n} \exp\left(-\frac{\lambda_a}{2\tau}\right) \cdot \prod_{a=1}^{n} \lambda_a^{(m-n-1)/2} \cdot \prod_{a<b}(\lambda_a - \lambda_b).$$

Hence, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{\sigma_i}{\sigma_i - \sigma_j}\right) \\
\propto\ & \int_{\lambda_1\geq\ldots\geq\lambda_n} \cdots \int \frac{\sqrt{\lambda_i}}{\sqrt{\lambda_i} - \sqrt{\lambda_j}} \cdot \prod_{a=1}^{n} \exp\left(-\frac{\lambda_a}{2\tau}\right) \cdot \prod_{a=1}^{n} \lambda_a^{(m-n-1)/2} \cdot \prod_{a<b}(\lambda_a - \lambda_b) d\boldsymbol{\lambda} \\
\leq\ & \int_{\lambda_1\geq\ldots\geq\lambda_n} \cdots \int 2\lambda_i \prod_{a=1}^{n} \exp\left(-\frac{\lambda_a}{2\tau}\right) \cdot \prod_{a=1}^{n} \lambda_a^{(m-n-1)/2} \cdot \prod_{\substack{a<b \\ (a,b)\neq(i,j)}} (\lambda_a - \lambda_b) d\boldsymbol{\lambda} \\
\leq\ & 2\int_{\lambda_1\geq\ldots\geq\lambda_n} \cdots \int \prod_{a=1}^{n} \exp\left(-\frac{\lambda_a}{2\tau}\right) \cdot \prod_{a=1}^{n} \lambda_a^{(m-n-1)/2} \cdot \prod_{a=1}^{n} \lambda_a^{n-a} d\boldsymbol{\lambda} \\
\leq\ & 2\prod_{a=1}^{n} \int_0^{+\infty} \exp\left(-\frac{\lambda_a}{2\tau}\right) \cdot \lambda_a^{(m-n-1)/2+n-a} d\lambda_a < \infty,
\end{aligned}$$

where the second inequality is simply due to $\lambda_a - \lambda_b \leq \lambda_a$ for $a < b$. Similarly, we can show $\mathbb{E}(1/(\sigma_i - \sigma_j)) < \infty$. Moreover,

$$\mathbb{E}(\sigma_i^2/(\sigma_i^2 - \sigma_j^2)) \leq \mathbb{E}(\sigma_i/(\sigma_i - \sigma_j)) < \infty.$$

When $M^* \neq \mathbf{0}$, we express $\sigma_i/(\sigma_i - \sigma_j)$ as a function of $M^* + \mathcal{E}$, denoted by $h(M^* + \mathcal{E})$. Then,

$$\mathbb{E}\left(\frac{\sigma_i}{\sigma_i - \sigma_j}\right) = \frac{1}{(2\pi)^{mn/2}\tau^{mn}} \int h(M^* + \mathcal{E}) \exp\left(-\frac{1}{2\tau^2}\|\mathcal{E}\|_F^2\right) d\mathcal{E}$$

$$= \frac{1}{(2\pi)^{mn/2}\tau^{mn}} \int h(\mathcal{E}) \exp\left(-\frac{1}{2\tau^2}\|\mathcal{E} - M^*\|_F^2\right) d\mathcal{E}$$
$$\leq \frac{1}{(2\pi)^{mn/2}\tau^{mn}} \cdot \exp(\|M^*\|_F^2/(2\tau^2)) \int h(\mathcal{E}) \exp\left(-\frac{1}{4\tau^2}\|\mathcal{E}\|_F^2\right) d\mathcal{E} \overset{(a)}{<} \infty,$$

where (a) is implied by the boundedness result for $M^* = \mathbf{0}$. Similar arguments work for the other two expectations. □

7.3. Proof of Corollary 2

According to Proposition 3 in Mazumder et al. [23], $S_\theta(Y)$ is Lipschitz continuous, which is sufficient for the regularity conditions to hold (see Lemma 3.2 in Candès et al. [4]). Since $s_\theta(\cdot)$ is Lipschitz, it is differentiable almost everywhere. Under the model (3.1), the singular values of Y have a multiplicity of one and are non-zero with probability one. It means that we only need to compute $\nabla \cdot S_\theta(Y)$ for a full rank matrix Y with singular values $\sigma_1 > \ldots > \sigma_n > 0$ at which $s_\theta(\cdot)$ is differentiable. A direct application of Corollary 1 gives the formula in (3.4).

7.4. Proof of Corollary 3

Denote the spectral regularized estimator in expression (3.2) with $P_\theta(\cdot)$ being the MC+ penalty by $S_{\sqrt{2\theta},\gamma}(Y)$. Specifically, $S_{\sqrt{2\theta},\gamma}(Y) = \sum_{i=1}^n g_{\sqrt{2\theta},\gamma}(\sigma_i)\boldsymbol{u}_i\boldsymbol{v}_i'$, where $g_{\sqrt{2\theta},\gamma}(\cdot)$ is a piecewise linear function defined on $[0, +\infty)$:

$$g_{\sqrt{2\theta},\gamma}(\sigma) = \begin{cases} 0 & \text{if } \sigma \leq \sqrt{2\theta} \\ \frac{\gamma(\sigma - \sqrt{2\theta})}{\gamma - 1} & \text{if } \sqrt{2\theta} < \sigma \leq \sqrt{2\theta}\gamma \\ \sigma & \text{if } \sigma > \sqrt{2\theta}\gamma. \end{cases}$$

Then it is easy to see that $S_{\sqrt{2\theta},\gamma}(Y) \to S_\theta(Y)$, as $\gamma \downarrow 1$. Hence we have,

$$\begin{aligned} |df(S_\theta(Y)) - df(S_{\sqrt{2\theta},\gamma}(Y))| &= \frac{1}{\tau^2}\Big|\sum_{ij} \mathbb{E}((S_\theta(Y))_{ij} - (S_{\sqrt{2\theta},\gamma}(Y))_{ij})\epsilon_{ij}\Big| \\ &\leq \frac{1}{\tau^2}\mathbb{E}\|S_\theta(Y) - S_{\sqrt{2\theta},\gamma}(Y)\|_F \cdot \|\mathcal{E}\|_F \\ &\leq \frac{1}{\tau^2}\mathbb{E}^{1/2}\|S_\theta(Y) - S_{\sqrt{2\theta},\gamma}(Y)\|_F^2 \cdot \mathbb{E}^{1/2}\|\mathcal{E}\|_F^2 \\ &\to 0 \quad \text{as } \gamma \downarrow 1, \end{aligned}$$

where the last line holds by using Dominated Convergence Theorem (DCT). We can apply DCT here because

$$\begin{aligned} &\|S_\theta(Y) - S_{\sqrt{2\theta},\gamma}(Y)\|_F^2 \leq (\|S_\theta(Y)\|_F + \|S_{\sqrt{2\theta},\gamma}(Y)\|_F)^2 \\ \leq &2(\|S_\theta(Y)\|_F^2 + \|S_{\sqrt{2\theta},\gamma}(Y)\|_F^2) \leq 4\|S_\theta(Y)\|_F^2 \leq 4\|Y\|_F^2. \end{aligned}$$

Therefore, we can calculate $df(S_\theta(Y))$ via the following limiting argument,

$$df(S_\theta(Y)) = \lim_{\gamma \downarrow 1} df(S_{\sqrt{2\theta},\gamma}(Y)).$$

When $\gamma > 1$, the operator $S_{\sqrt{2\theta},\gamma}(Y)$ satisfies the conditions in Corollary 2. Hence:

$$\begin{aligned} df(S_{\sqrt{2\theta},\gamma}(Y)) &= \sum_{i=1}^{n} \left(\frac{\gamma}{\gamma-1} P(\sqrt{2\theta} < \sigma_i \le \sqrt{2\theta}\gamma) + P(\sigma_i > \sqrt{2\theta}\gamma) \right) \\ &\quad + \mathbb{E}\left[\sum_{\substack{i \neq j \\ i,j=1}}^{n} \frac{\sigma_i g_{\sqrt{2\theta},\gamma}(\sigma_i) - \sigma_j g_{\sqrt{2\theta},\gamma}(\sigma_j)}{\sigma_i^2 - \sigma_j^2} \right] \end{aligned}$$

Now we calculate the limit of each term in the above equation. Let $F_{\sigma_i}(\cdot), f_{\sigma_i}(\cdot)$ be the cdf, pdf of σ_i respectively, and $f_{\sigma_i,\sigma_j}(\cdot,\cdot)$ the joint pdf of (σ_i, σ_j). It is straightforward to see $\lim_{\gamma\downarrow 1} P(\sigma_i > \sqrt{2\theta}\gamma) = P(\sigma_i > \sqrt{2\theta})$, and

$$\begin{aligned} \lim_{\gamma\downarrow 1} \frac{\gamma}{\gamma-1} P(\sqrt{2\theta} < \sigma_i \le \sqrt{2\theta}\gamma) &= \lim_{\gamma\downarrow 1} \sqrt{2\theta}\gamma \cdot \lim_{\gamma\downarrow 1} \frac{F_{\sigma_i}(\sqrt{2\theta}\gamma) - F_{\sigma_i}(\sqrt{2\theta})}{\sqrt{2\theta}(\gamma-1)} \\ &= \sqrt{2\theta} f_{\sigma_i}(\sqrt{2\theta}). \end{aligned}$$

Finally, we decompose $\mathbb{E}\left(\frac{\sigma_i g_{\sqrt{2\theta},\gamma}(\sigma_i) - \sigma_j g_{\sqrt{2\theta},\gamma}(\sigma_j)}{\sigma_i^2 - \sigma_j^2} \right)$ into 8 terms:

$$\begin{aligned} I_1 &= \mathbb{E}\mathbb{1}(\sigma_i \le \sqrt{2\theta}, \sqrt{2\theta} < \sigma_j \le \sqrt{2\theta}\gamma) \cdot \frac{\gamma\sigma_j(\sigma_j - \sqrt{2\theta})}{(\gamma-1)(\sigma_j^2 - \sigma_i^2)} \\ I_2 &= \mathbb{E}\mathbb{1}(\sigma_j \le \sqrt{2\theta}, \sqrt{2\theta} < \sigma_i \le \sqrt{2\theta}\gamma) \cdot \frac{\gamma\sigma_i(\sigma_i - \sqrt{2\theta})}{(\gamma-1)(\sigma_i^2 - \sigma_j^2)} \\ I_3 &= \mathbb{E}\mathbb{1}(\sigma_i \le \sqrt{2\theta}, \sigma_j > \sqrt{2\theta}\gamma) \cdot \frac{\sigma_j^2}{\sigma_j^2 - \sigma_i^2} \\ I_4 &= \mathbb{E}\mathbb{1}(\sigma_j \le \sqrt{2\theta}, \sigma_i > \sqrt{2\theta}\gamma) \cdot \frac{\sigma_i^2}{\sigma_i^2 - \sigma_j^2} \\ I_5 &= \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_i \le \sqrt{2\theta}\gamma, \sqrt{2\theta} < \sigma_j \le \sqrt{2\theta}\gamma) \cdot \frac{\gamma}{\gamma-1} \cdot \left(1 - \frac{\sqrt{2\theta}}{\sigma_i + \sigma_j}\right) \\ I_6 &= \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_i \le \sqrt{2\theta}\gamma, \sigma_j > \sqrt{2\theta}\gamma) \cdot \left(1 + \frac{1}{\gamma-1} \cdot \frac{\sigma_i^2 - \sqrt{2\theta}\gamma\sigma_i}{\sigma_i^2 - \sigma_j^2}\right) \\ I_7 &= \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_j \le \sqrt{2\theta}\gamma, \sigma_i > \sqrt{2\theta}\gamma) \cdot \left(1 + \frac{1}{\gamma-1} \cdot \frac{\sigma_j^2 - \sqrt{2\theta}\gamma\sigma_j}{\sigma_j^2 - \sigma_i^2}\right) \\ I_8 &= \mathbb{E}\mathbb{1}(\sigma_i > \sqrt{2\theta}\gamma, \sigma_j > \sqrt{2\theta}\gamma). \end{aligned}$$

We analyze each of the above terms individually. First, since $\mathbb{E}(1/|\sigma_j - \sigma_i|) < \infty$ by Lemma 3, it holds that

$$I_1 \le \mathbb{E}\mathbb{1}(\sigma_i \le \sqrt{2\theta}, \sqrt{2\theta} < \sigma_j \le \sqrt{2\theta}\gamma) \cdot \frac{\gamma\sigma_j(\sqrt{2\theta}\gamma - \sqrt{2\theta})}{(\gamma-1)(\sigma_j + \sigma_i)(\sigma_j - \sigma_i)}$$

$$\leq \mathbb{E}\mathbb{1}(\sigma_i \leq \sqrt{2\theta}, \sqrt{2\theta} < \sigma_j \leq \sqrt{2\theta}\gamma) \cdot \frac{\gamma\sqrt{2\theta}}{\sigma_j - \sigma_i}$$
$$\leq \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_j \leq \sqrt{2\theta}\gamma) \cdot \frac{\gamma\sqrt{2\theta}}{|\sigma_j - \sigma_i|} \to 0, \text{ as } \gamma \downarrow 1.$$

Similarly, we have $\lim_{\gamma\downarrow 1} I_2 = 0$. Because $\mathbb{E}(\sigma_j^2/|\sigma_j^2 - \sigma_i^2|) < \infty$ by Lemma 3, we have $\lim_{\gamma\downarrow 1} I_3 = \mathbb{E}\mathbb{1}(\sigma_i < \sqrt{2\theta}, \sigma_j > \sqrt{2\theta}) \cdot \sigma_j^2/(\sigma_j^2 - \sigma_i^2)$, and $\lim_{\gamma\downarrow 1} I_4 = \mathbb{E}\mathbb{1}(\sigma_j < \sqrt{2\theta}, \sigma_i > \sqrt{2\theta}) \cdot \sigma_i^2/(\sigma_i^2 - \sigma_j^2)$. Moreover,

$$I_5 \leq \frac{\gamma}{\gamma - 1} P(\sqrt{2\theta} < \sigma_i \leq \sqrt{2\theta}\gamma, \sqrt{2\theta} < \sigma_j \leq \sqrt{2\theta}\gamma) \to 0,$$

since $P(\sqrt{2\theta} < \sigma_i \leq \sqrt{2\theta}\gamma, \sqrt{2\theta} < \sigma_j \leq \sqrt{2\theta}\gamma) \sim (\gamma - 1)^2$. Also,

$$I_6 \leq \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_i \leq \sqrt{2\theta}\gamma, \sigma_j > \sqrt{2\theta}\gamma) \cdot \Big(1 + \frac{1}{\gamma - 1} \cdot \frac{\sigma_i(\sqrt{2\theta}\gamma - \sqrt{2\theta})}{(\sigma_j - \sigma_i)(\sigma_j + \sigma_i)}\Big)$$
$$\leq \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_i \leq \sqrt{2\theta}\gamma, \sigma_j > \sqrt{2\theta}\gamma) \cdot \Big(1 + \frac{\sqrt{2\theta}}{\sigma_j - \sigma_i}\Big)$$
$$\leq P(\sqrt{2\theta} < \sigma_i \leq \sqrt{2\theta}\gamma) + \mathbb{E}\mathbb{1}(\sqrt{2\theta} < \sigma_i \leq \sqrt{2\theta}\gamma) \cdot \frac{\sqrt{2\theta}}{|\sigma_j - \sigma_i|} \to 0.$$

Similarly, $\lim_{\gamma\downarrow 0} I_7 = 0$. Clearly, $\lim_{\gamma\downarrow 1} I_8 = P(\sigma_i > \sqrt{2\theta}, \sigma_j > \sqrt{2\theta})$. Collecting all the terms we analyzed so far leads to the *df* expression of rank regularized estimator.

7.5. Proof of Corollary 4

According to Lemmas 5–7 in [41], we can decompose the function $\eta_q(\sigma;\theta)$ over $[0,\infty)$ as follows:

$$\eta_q(\sigma;\theta) = \zeta_q(\sigma;\theta) + \xi_q(\sigma;\theta),$$

where $\zeta_q(\sigma;\theta) = [2(1-q)\theta]^{1/(2-q)} \cdot \mathbb{1}(\sigma > c_q\theta^{1/(2-q)})$, and $\xi_q(\sigma;\theta)$ is a Lipschitz continuous function. Let us define

$$\widetilde{S}_\theta(Y) = \sum_{i=1}^n \zeta_q(\sigma_i;\theta)\boldsymbol{u}_i\boldsymbol{v}_i' \quad \text{and} \quad \bar{S}_\theta(Y) = \sum_{i=1}^n \xi_q(\sigma_i;\theta)\boldsymbol{u}_i\boldsymbol{v}_i'.$$

By the definition of *df* in (1.2), we have

$$df(S_\theta(Y)) = df(\widetilde{S}_\theta(Y)) + df(\bar{S}_\theta(Y)).$$

Due to the Lipschitz continuity of $\xi_q(\sigma;\theta)$, we can use the same arguments as presented in the proof of Lemma 4 to conclude that $\bar{S}_\theta(Y)$ is Lipschitz continuous. Hence the formula (1.4) is applicable to $\bar{S}_\theta(Y)$. Its *df* can be computed

by the divergence formula in Corollary 1. Regarding the *df* of $\widetilde{S}_\theta(Y)$, similar to what we did in the proof of Corollary 3, we construct a sequence of approximations: $\widetilde{S}_{\theta,h}(Y) = \sum_{i=1}^n g_{\theta,h}(\sigma_i)\boldsymbol{u}_i\boldsymbol{v}_i'$, where $g_{\theta,h}$ is a piecewise linear function:

$$g_{\theta,h}(\sigma) = \begin{cases} 0 & \text{if } 0 \le \sigma < c_q\theta^{1/(2-q)} \\ \frac{[2(1-q)\theta]^{1/(2-q)}}{h}(\sigma - c_q\theta^{1/(2-q)}) & \text{if } c_q\theta^{1/(2-q)} \le \sigma \le c_q\theta^{1/(2-q)} + h \\ [2(1-q)\theta]^{1/(2-q)} & \text{if } \sigma > c_q\theta^{1/(2-q)} + h. \end{cases}$$

Because $\widetilde{S}_{\theta,h}(Y)$ is Lipschitz, we can compute $df(\widetilde{S}_{\theta,h}(Y))$ with the divergence formula in Corollary 1 and obtain $df(\widetilde{S}_\theta(Y))$ by letting $h \downarrow 0$. Since the calculations are very similar to those in the proof of Corollary 3, we do not repeat here. Adding up the *df* formulas of $\widetilde{S}_\theta(Y)$ and $\bar{S}_\theta(Y)$ finishes the proof.

7.6. *Proof of Corollary 5*

We consider the non-trivial case when $K < n$. The case $K = n$ can be directly verified. Before we go to the the main proof, we present two lemmas that will be used in the proof.

Lemma 4. *For any two matrices $Y_1, Y_2 \in \mathbb{R}^{m\times n}$, denote*

$$\mathcal{L} \triangleq \max\left\{\frac{\sigma_K(Y_1)}{\sigma_K(Y_1) - \sigma_{K+1}(Y_1)}, \frac{\sigma_K(Y_2)}{\sigma_K(Y_2) - \sigma_{K+1}(Y_2)}\right\}.$$

We then have

$$\|C_K(Y_1) - C_K(Y_2)\|_F \le \mathcal{L}\cdot\|Y_1 - Y_2\|_F.$$

Proof. Let $f_1(\sigma) = \sigma\mathbb{1}(\sigma \ge \sigma_K(Y_1))$ and $f_2(\sigma) = \sigma\mathbb{1}(\sigma \ge \sigma_K(Y_2))$. Then

$$\begin{aligned}
& \mathcal{L}^2\|Y_1 - Y_2\|_F^2 - \|C_K(Y_1) - C_K(Y_2)\|_F^2 \\
= \; & \sum_i [\mathcal{L}^2(\sigma_i^2(Y_1) + \sigma_i^2(Y_2)) - f_1^2(\sigma_i(Y_1)) - f_2^2(\sigma_i(Y_2))] \\
& -2\mathcal{L}^2\mathrm{tr}(Y_1'Y_2) + 2\mathrm{tr}(C_K'(Y_1)C_K(Y_2)) \\
= \; & \sum_i [\mathcal{L}^2(\sigma_i^2(Y_1) + \sigma_i^2(Y_2)) - f_1^2(\sigma_i(Y_1)) - f_2^2(\sigma_i(Y_2))] \\
& -2\mathrm{tr}[(\mathcal{L}Y_1 - C_K(Y_1))'(\mathcal{L}Y_2 - C_K(Y_2))] \\
& -2\mathrm{tr}[(\mathcal{L}Y_1 - C_K(Y_1))'C_K(Y_2)] - 2\mathrm{tr}[C_K(Y_1)'(\mathcal{L}Y_2 - C_K(Y_2))] \\
\overset{(a)}{\ge} \; & \sum_i [\mathcal{L}^2(\sigma_i^2(Y_1) + \sigma_i^2(Y_2)) - f_1^2(\sigma_i(Y_1)) - f_2^2(\sigma_i(Y_2))] \\
& -2\sum_i [\mathcal{L}\sigma_i(Y_1) - f_1(\sigma_i(Y_1))]\cdot[\mathcal{L}\sigma_i(Y_2) - f_2(\sigma_i(Y_2))] \\
& -2\sum_i [\mathcal{L}\sigma_i(Y_1) - f_1(\sigma_i(Y_1))]\cdot f_2(\sigma_i(Y_2))
\end{aligned}$$

$$
\begin{aligned}
& -2\sum_i f_1(\sigma_i(Y_1))\cdot[\mathcal{L}\sigma_i(Y_2) - f_2(\sigma_i(Y_2))] \\
\geq\ & \sum_i [\mathcal{L}^2(\sigma_i(Y_1)-\sigma_i(Y_2))^2 - (f_1(\sigma_i(Y_1)) - f_2(\sigma_i(Y_2)))^2] \\
\geq\ & 0,
\end{aligned}
$$

where, inequality (a) holds by (i) making use of Von Neumann's trace inequality [36]; and (ii) noting that the sequences $\{\mathcal{L}\sigma_i(Y_1) - f_1(\sigma_i(Y_1))\}_{i=1}^n$ and $\{\mathcal{L}\sigma_i(Y_2) - f_2(\sigma_i(Y_2))\}_{i=1}^n$ are descending[5] in i. □

Lemma 5. *Given any $Y \in \mathbb{R}^{m\times n}$, if $\sigma_K(Y) > \sigma_{K+1}(Y)$, then $C_K(Y)$ is directionally differentiable at Y and*

$$
\mathbb{E}_Z\left(\lim_{h\to 0+}\frac{([C_K(hZ+Y)]_{ij} - [C_K(Y)]_{ij})Z_{ij}}{h}\right) = \frac{\partial[C_K(Y)]_{ij}}{\partial Y_{ij}}, \tag{7.7}
$$

where the entries of Z follow i.i.d $N(0,1)$. Moreover,

$$
\sum_{ij}\frac{\partial[C_K(Y)]_{ij}}{\partial Y_{ij}} = (m+n-K)K + 2\sum_{i=1}^{K}\sum_{K+1}^{n}\frac{\sigma_j^2}{\sigma_i^2-\sigma_j^2}. \tag{7.8}
$$

Proof. Construct a function $v:\mathbb{R}^+\to\mathbb{R}^+$ as

$$
v(x) = \begin{cases} 0 & \text{if } x \leq \sigma_{K+1}(Y)+\Delta, \\ \frac{(\sigma_K(Y)-\Delta)(x-\sigma_{K+1}(Y)-\Delta)}{\sigma_K(Y)-\sigma_{K+1}(Y)-2\Delta} & \text{if } \sigma_{K+1}(Y)+\Delta < x \leq \sigma_K(Y)-\Delta, \\ x & \text{otherwise,} \end{cases}
$$

where Δ is a positive constant smaller than $(\sigma_K(Y)-\sigma_{K+1}(Y))/2$. It is straightforward to confirm that $v(0)=0$ and $v(\cdot)$ is differentiable at $\sigma_i(Y), 1\leq i\leq n$. Hence applying Theorem 1 gives

$$
\sum_{ij}\frac{\partial[S(Y;w)]_{ij}}{\partial Y_{ij}} = (m+n-K)K + 2\sum_{i=1}^{K}\sum_{K+1}^{n}\frac{\sigma_j^2}{\sigma_i^2-\sigma_j^2}.
$$

Note that since the singular values $\sigma_i(Y)$ are continuous, we know $C_K(\tilde{Y}) = S(\tilde{Y};w)$ for $\tilde{Y}$ in a small neighborhood of Y. This fact combined with the last equality proves (7.8). Regarding (7.7), since $v(\cdot)$ is differentiable at $\sigma_i(Y), 1\leq i\leq n$, we can combine Lemmas 1 and 2 (as we did in the proof of Theorem 1) to conclude that the directional derivative $\lim_{h\to 0+}([C_K(hZ+Y)]_{ij} - [C_K(Y)]_{ij})/h$ is linear in Z. Denoting the directional derivative by $D(Y)\in\mathbb{R}^{m\times n}$, we have:

$$
\mathbb{E}_Z\lim_{h\to 0+}\frac{([C_K(hZ+Y)]_{ij} - [C_K(Y)]_{ij})Z_{ij}}{h} = \mathbb{E}_Z[\mathrm{tr}(Z'D(Y))Z_{ij}] = [D(Y)]_{ij}
$$

[5] This follows from the choice of $\mathcal{L}$.

By the definition of $D(Y)$, we know that

$$[D(Y)]_{ij} = \lim_{h \to 0+} \frac{([C_K(he_{ij} + Y)]_{ij} - [C_K(Y)]_{ij})}{h} = \frac{\partial [C_K(Y)]_{ij}}{\partial Y_{ij}},$$

where, recall that e_{ij} is a matrix with its (i,j)th entry being one and the rest being zero. This completes the proof of (7.7). □

Consider a smoothed version of $C_K(Y)$, defined below

$$g_h(Y) \triangleq \mathbb{E}_Z[C_K(Y + hZ)],$$

where the elements of Z are i.i.d from $N(0,1)$, independent of Y; the expectation $\mathbb{E}_Z$ is taken only with respect to Z; and h is a positive constant. We will show that $df\,(g_h(Y))$ is a good approximation to $df\,(C_K(Y))$ i.e.,

$$\lim_{h \to 0+} df(g_h(Y)) = df(C_K(Y)). \tag{7.9}$$

To prove (7.9), by using the original definition of df, it suffices to show

$$\lim_{h \to 0+} \mathbb{E}([g_h(Y)]_{ij} Y_{ij}) = \mathbb{E}([C_K(Y)]_{ij} Y_{ij}), \quad \lim_{h \to 0+} \mathbb{E}([g_h(Y)]_{ij}) = \mathbb{E}([C_K(Y)]_{ij})$$

for all $1 \le i \le m, 1 \le j \le n$.

We prove the first equality above; the second one follows using a similar argument. First note that

$$\mathbb{E}([g_h(Y)]_{ij} Y_{ij}) - \mathbb{E}([C_K(Y)]_{ij} Y_{ij}) = \mathbb{E}(([C_K(Y + hZ)]_{ij} - [C_K(Y)]_{ij}) Y_{ij})$$

Since $\|C_K(Y)\|_F \le \|Y\|_F$ for any $Y \in \mathbb{R}^{m \times n}$, we have for small h

$$|([C_K(Y + hZ)]_{ij} - [C_K(Y)]_{ij}) Y_{ij}| \le (\|Z\|_F + 2\|Y\|_F)\|Y\|_F.$$

Hence we can use the Dominated Convergence Theorem (DCT) to conclude

$$\begin{aligned} &\lim_{h \to 0+} \mathbb{E}(([C_K(Y + hZ)]_{ij} - [C_K(Y)]_{ij}) Y_{ij}) \\ &= \mathbb{E} \lim_{h \to 0+} (([C_K(Y + hZ)]_{ij} - [C_K(Y)]_{ij}) Y_{ij}) \overset{(b)}{=} 0. \end{aligned}$$

To derive (b) we have used the fact that $C_K(Y)$ is directionally differentiable from Lemma 5. Based on (7.9), we can compute $df(C_K(Y))$ by first calculating $df(g_h(Y))$ and then use a limiting argument with $h \downarrow 0$. Since $g_h(Y)$ is differentiable, it is straightforward to get

$$\begin{aligned} df(g_h(Y)) &= \sum_{ij} \mathbb{E}\Big(\frac{\partial [g_h(Y)]_{ij}}{\partial Y_{ij}}\Big) = \sum_{ij} \mathbb{E}\Big(\frac{[C_K(hZ + Y)]_{ij} Z_{ij}}{h}\Big) \\ &\overset{(c)}{=} \sum_{ij} \mathbb{E}\Big(\frac{([C_K(hZ + Y)]_{ij} - [C_K(Y)]_{ij}) Z_{ij}}{h}\Big), \end{aligned} \tag{7.10}$$

where (c) holds because Z is independent of Y and has zero mean. We seek to compute the following limits:

$$\lim_{h\to 0+} \mathbb{E}\Big(\frac{([C_K(hZ+Y)]_{ij}-[C_K(Y)]_{ij})Z_{ij}}{h}\Big) = \lim_{h\to 0+} \mathbb{E}_Z\ J(Z,h) \tag{7.11}$$

where,

$$J(Z,h) \triangleq \int \frac{([C_K(hZ+Y)]_{ij}-[C_K(Y)]_{ij})Z_{ij}}{h} \frac{1}{(\sqrt{2\pi}\tau)^{mn}} \exp\Big(\frac{\|Y-M^*\|_F^2}{-2\tau^2}\Big) dY.$$

According to Lemma 4, we can obtain

$$\begin{aligned} |J(Z,h)| \le \|Z\|_F |Z_{ij}| \cdot \mathbb{E}_Y \frac{\sigma_K(Y)}{\sigma_K(Y)-\sigma_{K+1}(Y)} \quad + \\ \|Z\|_F |Z_{ij}| \cdot \mathbb{E}_Y \frac{\sigma_K(hZ+Y)}{\sigma_K(hZ+Y)-\sigma_{K+1}(hZ+Y)}. \end{aligned} \tag{7.12}$$

Moreover, a simple change of variable gives us

$$\begin{aligned} &\mathbb{E}_Y \frac{\sigma_K(hZ+Y)}{\sigma_K(hZ+Y)-\sigma_{K+1}(hZ+Y)} \\ &= \int \frac{\sigma_K(Y)}{\sigma_K(Y)-\sigma_{K+1}(Y)} \frac{1}{(\sqrt{2\pi}\tau)^{mn}} \exp\Big(\frac{\|Y-hZ-M^*\|_F^2}{-2\tau^2}\Big) dY \\ &\le \frac{1}{(\sqrt{2\pi}\tau)^{mn}} \exp\Big(\frac{\|hZ+M^*\|_F^2}{2\tau^2}\Big) \int \frac{\sigma_K(Y)}{\sigma_K(Y)-\sigma_{K+1}(Y)} \exp\Big(\frac{\|Y\|_F^2}{-4\tau^2}\Big) dY \\ &\le \frac{1}{(\sqrt{2\pi}\tau)^{mn}} \exp\Big(\frac{\|M^*\|_F^2}{\tau^2}\Big) \cdot \exp\Big(\frac{h^2\|Z\|_F^2}{\tau^2}\Big) \cdot \\ &\quad \int \frac{\sigma_K(Y)}{\sigma_K(Y)-\sigma_{K+1}(Y)} \exp\Big(\frac{\|Y\|_F^2}{-4\tau^2}\Big) dY \end{aligned} \tag{7.13}$$

Combining Lemma 3 part (1) with (7.12) and (7.13), we can conclude that for sufficiently small h, there exists an upper bound on $J(Z,h)$ that is independent of h and is integrable. We thus can employ DCT to get

$$\lim_{h\to 0+} \mathbb{E}_Z[J(Z,h)] = \mathbb{E}_Z \lim_{h\to 0+} [J(Z,h)]. \tag{7.14}$$

We next focus on calculating $\lim_{h\to 0+}[J(Z,h)]$. We decompose $J(Z,h)$ into two terms:

$$\begin{aligned} J(Z,h) = \mathbb{E}_Y \Bigg[&\frac{([C_K(hZ+Y)]_{ij}-[C_K(Y)]_{ij})Z_{ij}}{h} \cdot \\ &\mathbb{1}(\sigma_K(Y) \ge h^{2/3}, \sigma_K(Y)-\sigma_{K+1}(Y) \ge h^{2/3}) \Bigg] + \\ \mathbb{E}_Y \Bigg[&\frac{([C_K(hZ+Y)]_{ij}-[C_K(Y)]_{ij})Z_{ij}}{h} \cdot \end{aligned} \tag{7.15}$$

$$\mathbb{1}(\sigma_K(Y) \leq h^{2/3} \text{ or } \sigma_K(Y) - \sigma_{K+1}(Y) \leq h^{2/3})\Big].$$

Denoting the two terms by $H_1(Y, Z, h)$ and $H_2(Y, Z, h)$ respectively, we analyze them separately. Regarding $H_1(Y, Z, h)$, first note that according to Weyl's inequality [30], we know

$$|\sigma_i(hZ + Y) - \sigma_i(Y)| \leq h\|Z\|_F, \quad 1 \leq i \leq n$$

Therefore, on the event $\{\sigma_K(Y) \geq h^{2/3}, \sigma_K(Y) - \sigma_{K+1}(Y) \geq h^{2/3}\}$, when h is sufficiently small, we have:

$$\frac{\sigma_K(hZ + Y)}{\sigma_K(hZ + Y) - \sigma_{K+1}(hZ + Y)} \leq \frac{4\sigma_K(Y)}{\sigma_K(Y) - \sigma_{K+1}(Y)}.$$

We can then employ Lemma 4 to obtain,

$$|H_1(Y, Z, h)| \leq \|Z\|_F |Z_{ij}| \frac{4\sigma_K(Y)}{\sigma_K(Y) - \sigma_{K+1}(Y)}.$$

This enables us to apply DCT to derive

$$\mathbb{E}_Z \lim_{h\to 0+} \mathbb{E}_Y H_1(Y, Z, h) = \mathbb{E} \lim_{h\to 0+} H_1(Y, Z, h) \overset{(d)}{=} \mathbb{E}_Y \frac{\partial [C_K(Y)]_{ij}}{\partial Y_{ij}}, \tag{7.16}$$

where (d) is due to Lemma 5. For the term $H_2(Y, Z, h)$, we have

$$\begin{aligned}
&|\mathbb{E}_Y H_2(Y, Z, h)| \leq \\
&\mathbb{E}_Y[|Z_{ij}| \cdot |(\|Z\|_F + 2\|Y\|_F/h) \cdot \mathbb{1}(\sigma_K(Y) \leq h^{2/3} \text{ or } \sigma_K(Y) - \sigma_{K+1}(Y) \leq h^{2/3})] \\
&\overset{(e)}{\leq} |Z_{ij}| \cdot \|Z\|_F \cdot P(\sigma_K(Y) - \sigma_{K+1}(Y) \leq h^{2/3}) + \\
&\quad 2|Z_{ij}| \cdot (\mathbb{E}\|Y\|_F^7)^{1/7} \cdot \Big(\frac{P(\sigma_K(Y) - \sigma_{K+1}(Y) \leq h^{2/3})}{h^{7/6}}\Big)^{6/7}.
\end{aligned} \tag{7.17}$$

We have used Hölder's inequality to derive (e). Clearly the first term of the upper bound above vanishes as $h \to 0+$. We now show the second term goes to zero as well. For simplicity, we only show it for $M^* = 0$. The general case $M^* \neq 0$ can be proved by the same arguments as presented in the proof of Lemma 3. We hence skip it here. Similar to the proof in Lemma 3, let $\lambda_i = \sigma_i^2(Y), 1 \leq i \leq n$ and denote the joint distribution of $(\lambda_1, \ldots, \lambda_n)$ by $f(\lambda_1, \ldots, \lambda_n)$. We can then rewrite

$$\begin{aligned}
&P(\sigma_K(Y) - \sigma_{K+1}(Y) \leq h^{2/3}) \\
&= \int_{\lambda_1 \geq \ldots \geq \lambda_n \geq 0} \cdots \int f(\lambda_1, \ldots, \lambda_n) \cdot \mathbb{1}(\sqrt{\lambda_K} - \sqrt{\lambda_{K+1}} \leq h^{2/3}) d\boldsymbol{\lambda} \propto \\
&\int_{\lambda_1 \geq \ldots \geq \lambda_n \geq 0} \cdots \int \mathbb{1}(\sqrt{\lambda_K} - \sqrt{\lambda_{K+1}} \leq h^{2/3}) \prod_{a=1}^n e^{\frac{\lambda_a}{-2\tau}} \cdot \prod_{a=1}^n \lambda_a^{(m-n-1)/2} \cdot \prod_{a<b} (\lambda_a - \lambda_b) d\boldsymbol{\lambda}
\end{aligned}$$

$$\overset{(f)}{\leq} \int_{\lambda_1 \geq \ldots \geq \lambda_n \geq 0} \cdots \int \mathbb{1}(\sqrt{\lambda_K} - \sqrt{\lambda_{K+1}} \leq h^{2/3}) \prod_{a=1}^{n} e^{\frac{\lambda_a}{-2\tau}} \cdot \prod_{a \neq K}^{n} \lambda_a^{(m+n-1)/2-a} \cdot \lambda_K^{(m+n-3)/2-K} (\lambda_K - \lambda_{K+1}) d\boldsymbol{\lambda}$$

$$\overset{(g)}{\leq} \iint_{0 \leq \lambda_{K+1} \leq \lambda_K} \mathbb{1}(\sqrt{\lambda_K} - \sqrt{\lambda_{K+1}} \leq h^{2/3}) e^{\frac{\lambda_K + \lambda_{K+1}}{-2\tau}} (\lambda_K \lambda_{K+1})^{(m+n-3)/2-K} \cdot (\lambda_K - \lambda_{K+1}) d\lambda_K d\lambda_{K+1}$$

$$\cdot \Big[\prod_{a \neq K, K+1}^{n} \int_0^{\infty} e^{\frac{\lambda_a}{-2\tau}} \lambda_a^{(m-n-1)/2+n-a} d\lambda_a \Big], \tag{7.18}$$

where, inequality (f) is obtained by using $\lambda_a - \lambda_b \leq \lambda_a$, for $a < b$; and inequality (g) holds simply because we enlarge the set over which the integration is performed. We easily see that the second term on the right hand side of the last inequality is finite and independent of h. We denote the first term by $Q(h)$. For this term, by using $\lambda_K - \lambda_{K+1} \leq 2\sqrt{\lambda_K}(\sqrt{\lambda_K} - \sqrt{\lambda_{K+1}})$, we have

$$Q(h) \leq 2h^{2/3} \iint_{0 \leq \lambda_{K+1} \leq \lambda_K} \mathbb{1}(\sqrt{\lambda_K} - \sqrt{\lambda_{K+1}} \leq h^{2/3}) e^{\frac{\lambda_K + \lambda_{K+1}}{-2\tau}} \lambda_K^{(m+n)/2-K-1} \cdot \lambda_{K+1}^{(m+n-3)/2-K} d\lambda_K d\lambda_{K+1}$$

$$= 2h^{2/3} \int_0^{\infty} \Big[\int_{\lambda_{K+1}}^{(\sqrt{\lambda_{K+1}} + h^{2/3})^2} e^{\frac{\lambda_K}{-2\tau}} \lambda_K^{(m-n)/2+n-K-1} d\lambda_K \Big] \cdot e^{\frac{\lambda_{K+1}}{-2\tau}} \lambda_{K+1}^{(m-n-1)/2+n-K-1} d\lambda_{K+1} \tag{7.19}$$

$$\overset{(h)}{=} O(h^{4/3}), \tag{7.20}$$

where (h) can be derived by using mean value theorem for the integral appearing in (7.19). Combining (7.17), (7.18) and (7.20) together gives us

$$\lim_{h \to 0+} \mathbb{E}_Y H_2(Y, Z, h) = 0. \tag{7.21}$$

Collecting the results from (7.9), (7.10), (7.11), (7.14), (7.15), (7.16) and (7.21), we can finally conclude

$$df(C_K(Y)) \quad = \quad \mathbb{E} \sum_{ij} \frac{\partial [C_K(Y)]_{ij}}{\partial Y_{ij}}.$$

A direct application of Equation (7.8) from Lemma 5 completes the proof.

7.7. *Proof of Corollary 6*

Denote the compact SVD of X by $X = U\Sigma V'$. We construct an ancillary matrix $Q = U'Y$, which is the response matrix of the following additive model,

$$Q = \tilde{M} + \tilde{\mathcal{E}}, \tag{7.22}$$

where $\tilde{M} = U'XM^*, \tilde{\mathcal{E}} = U'\mathcal{E}$. Due to the orthogonality of the columns of U, the entries of $\tilde{\mathcal{E}}$, i.e., $\tilde{\epsilon}_{ij} \overset{iid}{\sim} N(0, \tau^2)$. We now relate $df(M_K(Y))$ under the model (4.1) to $df(C_K(Q))$ under the model (7.22). A key observation is

$$XM_K(Y) = C_K(UU'Y) = UC_K(Q).$$

It follows that

$$\mathbb{E}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}(XM_K(Y))_{ij}\epsilon_{ij}\right) = \mathbb{E}\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{r}u_{ik}(C_K(Q))_{kj}\epsilon_{ij}\right),$$
$$\mathbb{E}\left(\sum_{a=1}^{r}\sum_{b=1}^{n}(C_K(Q))_{ab}\tilde{\epsilon}_{ab}\right) = \mathbb{E}\left(\sum_{a=1}^{r}\sum_{b=1}^{n}\sum_{l=1}^{m}(C_K(Q))_{ab}u_{la}\epsilon_{lb}\right),$$

where u_{ik} is the (i,k)th entry of U and $(C_K(Q))_{ab}$ is the (a,b)th element of $C_K(Q)$. Arranging the notation $a = k, b = j, l = i$, we thus obtain $df(M_K(Y)) = df(C_K(Q))$. Given that Q and $\hat{Y}$ share the same singular values, a direct use of Corollary 5 for $C_K(Q)$ gives us the df formula for $M_K(Y)$.

7.8. *Proof of Corollaries 7 and 8*

Observe that

$$XRM_\theta(Y) = S_\theta(UU'Y) = US_\theta(U'Y).$$

Thus we can use the same arguments as in the proof of Corollary 6 to obtain

$$df(RM_\theta(Y)) = df(S_\theta(U'Y)).$$

Then we use Corollaries 2 and 4 to complete the proof of Corollaries 7 and 8, respectively.

7.9. *Stein's unbiased risk estimate*

Proposition. *[29, 9, 17] Suppose* $\mathbf{y} \sim N(\boldsymbol{\mu}, \tau^2\mathbf{I}_n)$, $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^n$ *is weakly differentiable, and* $\mathbb{E}|y_i h_i(\mathbf{y})| + \mathbb{E}|\frac{\partial h_i(\mathbf{y})}{\partial y_i}| < \infty$ *for* $i = 1, \ldots, n$. *Then*

$$df(\boldsymbol{h}(\mathbf{y})) = \mathbb{E}\Big(\sum_{i=1}^{n}\partial h_i(\mathbf{y})/\partial y_i\Big),$$
$$\mathbb{E}\|\boldsymbol{h}(\mathbf{y}) - \boldsymbol{\mu}\|_2^2 = \mathbb{E}\Big[-\tau^2 n + \|\boldsymbol{h}(\mathbf{y}) - \mathbf{y}\|_2^2 + 2\tau^2 \cdot \sum_{i=1}^{n}\partial h_i(\mathbf{y})/\partial y_i\Big].$$

A function $g : \mathbb{R}^n \to \mathbb{R}$ *is said to be weakly differentiable if there exist functions* $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, n$, *such that for all compactly supported and infinitely differentiable functions* φ,

$$\int \varphi(\boldsymbol{z})h(\boldsymbol{z})d\boldsymbol{z} = -\int \frac{\partial\varphi(\boldsymbol{z})}{\partial z_i}g(\boldsymbol{z})d\boldsymbol{z}.$$

Acknowledgements

This work started as a part of a class project when both the authors were at Columbia University. Rahul Mazumder's research was partially supported by ONR-N000141512342, ONR-N000141812298 (YIP) and NSF-IIS1718258.

References

[1] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics* 327–351. MR0042664

[2] Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics* 1282–1309. MR2816355

[3] Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, **9** 717–772. MR2565240

[4] Candès, E. J., Sing-Long, C. A. and Trzasko, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, **61** 4643–4657. MR3105401

[5] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90** 1200–1224. MR1379464

[6] Drusvyatskiy, D. and Kempton, C. (2015). Variational analysis of spectral functions simplified. *arXiv preprint arXiv:1506.05170*.

[7] Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1** 211–218.

[8] Edelman, A. (2005). Matrix jacobians with wedge products.

[9] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, **32** 407–499. MR2060166

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96** 1348–1360. MR1946581

[11] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35** 109–135.

[12] Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics Springer, Berlin. MR2722294

[13] Gao, H.-Y. and Bruce, A. G. (1997). Waveshrink with firm shrinkage. *Statistica Sinica* 855–874. MR1488646

[14] Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, vol. 3. JHU Press. MR3024913

[15] Hansen, N. R. (2018). On stein's unbiased risk estimate for reduced rank estimators. *Statistics & Probability Letters*, **135** 76–82. MR3758265

[16] Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, **5** 248–264. MR0373179

[17] Johnstone, I. M. (2017). Gaussian estimation: Sequence and wavelet models. *Manuscript, August*.
[18] Lewis, A. S. and Sendov, H. S. (2005). Nonsmooth analysis of singular values. part i: Theory. *Set-Valued Analysis*, **13** 213–241. MR2162512
[19] Lewis, A. S. and Sendov, H. S. (2005). Nonsmooth analysis of singular values. part ii: Applications. *Set-Valued Analysis*, **13** 243–264. MR2162513
[20] Mallows, C. L. (1973). Some comments on c p. *Technometrics*, **15** 661–675.
[21] Mazumder, R., Friedman, J. H. and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, **106**. MR2894769
[22] Mazumder, R., Hastie, T. and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, **11** 2287–2322. MR2719857
[23] Mazumder, R., Saldana, D. F. and Weng, H. (2018). Matrix completion with nonconvex regularization: Spectral operators and scalable algorithms. *arXiv preprint arXiv:1801.08227*.
[24] Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, vol. 197. John Wiley & Sons. MR0652932
[25] Mukherjee, A., Chen, K., Wang, N. and Zhu, J. (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, **102** 457–477. MR3371016
[26] Pang, J.-S., Sun, D. and Sun, J. (2003). Semismooth homeomorphisms and strong stability of semidefinite and lorentz complementarity problems. *Mathematics of Operations Research*, **28** 39–63. MR1961266
[27] Papadopoulo, T. and Lourakis, M. I. (2000). Estimating the jacobian of the singular value decomposition: Theory and applications. In *Computer Vision-ECCV 2000*. Springer, 554–570.
[28] Shapiro, A. (2002). On differentiability of symmetric matrix valued functions. *Georgia Institute of Technology, available at* http://www.optimization-online.org/DB_FILE/2002/07/499.pdf.
[29] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* 1135–1151. MR0630098
[30] Stewart, G. W. (1998). Perturbation theory for the singular value decomposition. *Technical Report*.
[31] Sun, D. and Sun, J. (2002). Semismooth matrix-valued functions. *Mathematics of Operations Research*, **27** 150–169. MR1886224
[32] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288. MR1379242
[33] Tibshirani, R. J. (2015). Degrees of freedom and model search. *Statistica Sinica* 1265–1296. MR3410308
[34] Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, **40** 1198–1232. MR2985948
[35] Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, vol. 136. Springer Science & Business Media.

MR1719704

[36] Von Neumann, J. (1937). Some matrix-inequalities and metrization of matric-space. *Tomsk. Univ. Rev.*, **1** 286–300.

[37] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93** 120–131. MR1614596

[38] Yuan, M. (2011). Degrees of freedom in low rank matrix estimation. *Georgia Institute of Technology, available at* *http://pages.stat.wisc.edu/~myuan/papers/matcp.pdf*.

[39] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69** 329–346. MR2323756

[40] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 894–942. MR2604701

[41] Zheng, L., Maleki, A., Weng, H., Wang, X. and Long, T. (2017). Does ℓ_p-minimization outperform ℓ_1-minimization? *IEEE Transactions on Information Theory*, **63** 6896–6935. MR3724407

[42] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, **35** 2173–2192. MR2363967