Degrees of Freedom in Low Rank Matrix Estimation

Ming Yuan[†]

Georgia Institute of Technology

(November 18, 2011)

Abstract

The objective of this paper is to quantify the complexity of rank and nuclear norm constrained methods for low rank matrix estimation problems. Specifically, we derive analytic forms of the degrees of freedom for these types of estimators in several common settings. These results provide efficient ways of comparing different estimators and eliciting tuning parameters. Moreover, our analyses reveal new insights on the behavior of these low rank matrix estimators. These observations are of great theoretical and practical importance. In particular, they suggest that, contrary to conventional wisdom, for rank constrained estimators the total number of free parameters underestimates the degrees of freedom, whereas for nuclear norm penalization, it overestimates the degrees of freedom. In addition, when using most model selection criteria to choose the tuning parameter for nuclear norm penalization, it oftentimes suffices to entertain a finite number of candidates as opposed to a continuum of choices. Numerical examples are also presented to illustrate the practical implications of our results.

Keywords: Degrees of freedom, low rank matrix approximation, matrix completion, model selection, multivariate linear regression, nuclear norm penalization, reduced rank regression, singular value decomposition, Stein's unbiased risk estimator.

[†] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332. This research was supported in part by NSF Career Award DMS-0846234.

1 Introduction

The problem of low-rank matrix estimation naturally arises in a number of statistical and machine learning tasks. Prominent examples include multivariate linear regression, factor analysis, relational learning, multi-task learning, and matrix completion among many others. Numerous estimation methods have been developed in these contexts. Two of the most popular approaches are the rank constrained estimator, also called as reduced rank regression in the context of multivariate linear regression; and the nuclear norm regularized estimator, oftentimes referred to as matrix Lasso. A challenge common to both methods is how to effectively choose the tuning parameter and more fundamentally, how to assess the accuracy of an estimator having constructed it from a set of observations. It is well known that this goal cannot be achieved by simply measuring the estimator's fidelity to the same data on which it is computed, which inevitably leads to overoptimism about its performance (see, e.g., Efron, 1983; 1986). This issue is usually addressed by recalibrating the goodness of fit of an estimating procedure according to its complexity, a familiar idea behind the likes of Akaike information criterion (Akaike, 1973), Mallow's C_p (Mallows, 1973), Bayesian information criterion (Schwartz, 1978), generalized cross-validation (Craven and Wahba, 1979), Stein's unbiased risk estimate (Stein, 1981), and risk inflation criterion (Foster and George, 1994), to name a few. A recurring notion among these techniques is the so-called degrees of freedom which measures the complexity of an estimating procedure.

The importance of degrees of freedom in model assessment has long been recognized. Donoho and Johnstone (1995) derived an unbiased estimator of the degrees of freedom for soft thresholding and used it to find the optimal shrinkage factor in a wavelet denoising setting. More recently Efron (2004) showed that when using the correct degrees of freedom, a C_p type of statistic provides unbiased estimator of the prediction error, and in many cases offers substantial improvement over alternative techniques such as cross-validation. The significance of degrees of freedom has also been noted by Ye (1998), Shen and Ye (2002), Shen, Huang and Ye (2004), Zou, Hastie and Tibshirani (2007), among others.

The concept of degrees of freedom is most well-understood for linear estimators in the usual regression setting where it is identified with the trace of the so-called "hat" matrix (see, e.g., Hastie and Tibshirani, 1990). In particular, when considering the classical linear regression or the analysis of variance (ANOVA), it is often associated with the number

of variables in the model. In general, degrees of freedom can be rigorously defined in the framework of Stein's unbiased risk estimate (see, e.g., Ye, 1998; Efron, 2004). Its interpretation, however, is unclear in the context of low rank matrix estimation problems where the estimators are highly nonlinear in nature. Consider, for example, the popular reduced rank regression for multivariate linear regression. The number of free parameters in specifying a low rank matrix is often used as the degrees of freedom in this case (see, e.g., Reinsel and Velu, 1994). Although intuitive, it remains an open problem to what extent this appropriately measures the complexity of the rank constrained estimator. The main goal of this paper is to address such issues in a large class of low rank matrix estimation problems including among others the noisy singular value decomposition, reduced rank regression, and the more recently developed nuclear norm penalization.

Low rank matrix estimation methods often draw comparison with approaches for variable selection in the classical linear regression. In particular, the rank constrained estimator and the nuclear norm regularized estimator are reminiscent of the subset selection and the Lasso for linear regression whose degrees of freedom can be conveniently interpreted as the number of variables (Stein, 1981; Zou, Hastie and Tibshirani, 2006). This connection seemingly vindicates the number of free parameters as the degrees of freedom for their matrix analogues. However, as we show here, the number of free parameters incorrectly measures the complexity of either estimator. For the rank constrained estimator, the number of free parameters underestimates the degrees of freedom, whereas for the nuclear normal penalization, it overestimates the degrees of freedom. Furthermore, we provide explicit bias correction terms to rectify such a problem. Unlike the earlier developments where the degrees of freedom are estimated only through computationally intensive numerical methods such as data-perturbation or resampling procedures, we derive easily computable analytic forms of the degrees of freedom for several commonly used estimation procedures. In addition to the reduction of computational cost, our results reveal interesting insights about the behavior of these methods. These insights are of great theoretical and practical importance. For example, they suggest that when eliciting the tuning parameters for nuclear norm penalization, it may suffice to entertain a finite number of candidates rather than entertaining a continuum of choices.

The rest of the paper is organized as follows. We start in the next section with a canonical

low rank approximation/estimation problem where the goal is to estimate a low rank Gaussian mean matrix. Examples of such a problem include singular value decomposition with noise (Hoff, 2006), the analysis of relational data (Harshman et al., 1982), biplot (Gabriel, 1971; Gower and Hand, 1996), and reduced-rank interaction models for factorial designs (Gabriel 1978, 1998), among many others. We propose closed-form degrees of freedom estimators for both the rank constrained and nuclear norm penalized estimators.

In Section 3, we consider a couple related low rank matrix estimation problems, namely reduced rank regression for the multivariate linear regression (see, e.g., Reinsel and Velu, 1998) and nuclear norm penalization for matrix completion under uniform sampling at random (see, e.g., Koltchinskii, Lounici and Tsybakov, 2011). We show that analytic forms of unbiased degrees of freedom estimators can also be derived in these settings. Numerical experiments are reported in Section 4 to demonstrate the efficacy of the proposed estimators and their practical merits. All technical derivations are relegated to Section 5.

2 Canonical Low Rank Matrix Estimation

Many low rank matrix estimation problem can be formulated in the canonical form where the goal is to estimate a $m_1 \times m_2$ matrix M given a noisy observation $Y = M + E$, where the noise matrix E follows a matrix norm distribution $\mathrm{N}(0, \tau^2 \mathbf{I}_{m_1} \otimes \mathbf{I}_{m_2})$. Without loss of generality, we shall assume that $m_1 \leq m_2$ hereafter.

2.1 Degrees of Freedom

Let $\hat{M}$ be an estimate of M based on Y. Its degrees of freedom can be motivated as follows. Consider assessing the performance of $\hat{M}$ by $\|\hat{M} - M\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ stands for the usual matrix Frobenius or Hilbert-Schmidt norm. Observe that

$$\begin{aligned} \|\hat{M} - M\|_{\mathrm{F}}^2 &= \|\hat{M} - (Y - E)\|_{\mathrm{F}}^2 \\ &= \|\hat{M} - Y\|_{\mathrm{F}}^2 + 2\langle \hat{M} - Y, E\rangle + \|E\|_{\mathrm{F}}^2 \\ &= \|\hat{M} - Y\|_{\mathrm{F}}^2 + 2\langle \hat{M}, E\rangle + (\text{terms not depending on } \hat{M}), \end{aligned}$$

where $\langle \cdot, \cdot \rangle$ is the inner product associated with Frobenius norm, i.e., $\langle A, B\rangle = \mathrm{trace}(A^{\mathsf{T}} B)$. It is clear that the first term measures the goodness of fit of $\hat{M}$ to the observations Y. The

second term can then be interpreted as the cost of the estimating procedure leading to the following definition of degrees of freedom

$$\mathrm{df}(\hat{M}) = \frac{1}{\tau^2}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\mathrm{cov}(\hat{M}_{ij}, E_{ij}).$$

See Ye (1998) and Efron (2004) for further discussions. Once the degrees of freedom are defined, various performance evaluation criteria can be constructed for $\hat{M}$. In particular, the previous derivation suggests the following C_p type statistic:

$$C_p(\hat{M}) = \|\hat{M} - Y\|_{\mathrm{F}}^2 + 2\tau^2\mathrm{df}(\hat{M}).$$

Another popular alternative which we shall also focus on is the so-called generalized cross validation:

$$\mathrm{GCV}(\hat{M}) = \|\hat{M} - Y\|_{\mathrm{F}}^2/\{m_1m_2 - \mathrm{df}(\hat{M})\}^2.$$

Compared with other criteria, GCV has the advantage of not requiring τ^2 which is typically not known apriori and needs to be estimated from the data.

Generally speaking, the degrees of freedom as defined above are not directly computable. Stein (1981) solves this problem by constructing an unbiased estimator for it. In our context, his results indicate that

$$\mathrm{df}(\hat{M}) = \mathbb{E}\left(\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\partial\hat{M}_{ij}}{\partial E_{ij}}\right),$$

and suggest the following unbiased estimator of degrees of freedom:

$$\widehat{\mathrm{df}}^{\mathrm{S}}(\hat{M}) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\frac{\partial\hat{M}_{ij}}{\partial E_{ij}}.$$

However, with few exceptions, it is typically difficult to derive analytical expressions of $\widehat{\mathrm{df}}^{\mathrm{S}}(\hat{M})$. One often has to resort to numerical methods such as data perturbation and resampling techniques to compute it. These approaches, however, can be computationally prohibitive in large scale problems. It is therefore of great interests to derive, if possible at all, rigorous analytical results on the degrees of freedom. We now show that this indeed is possible for two of the most common low rank matrix estimators – rank regularized and nuclear norm regularized estimators.

2.2 Rank Regularized Estimator and Its Degrees of Freedom

We begin with rank constrained estimator. In the current context, it is given by:

$$\hat{M}^{\text{rank}}(K) = \underset{A \in \mathbb{R}^{m_1 \times m_2} : \text{rank}(A) \leq K}{\text{argmin}} \|A - Y\|_{\text{F}}^2;$$

where $K \in \{1, \ldots, m_1\}$ is a tuning parameter. The Eckart-Young Theorem shows that $\hat{M}^{\text{rank}}(K)$ is related to the singular value decomposition of Y and can be computed explicitly. More specifically, let $Y = U\Sigma V^{\mathsf{T}}$ be its singular value decomposition, i.e., Σ is a diagonal matrix with diagonal entries $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m_1} \geq 0$ and the column vectors of U and V are orthonormal. The reduced rank estimator $\hat{M}^{\text{rank}}(K)$ is well defined whenever $\sigma_K > \sigma_{K+1}$, which holds true with probability one. Moreover,

$$\hat{M}^{\text{rank}}(K) = \sum_{k=1}^{K} \sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}},$$

where $\mathbf{u}_k$ and $\mathbf{v}_k$ are the kth columns of U and V respectively.

Theorem 1 *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_K > \sigma_{K+1} \geq \sigma_{m_1}$ be the singular values of $Y = M + E$ where $E \sim \text{N}(0, \tau^2 \mathbf{I}_{m_1} \otimes \mathbf{I}_{m_2})$. Then an unbiased estimator of the degrees of freedom for $\hat{M}^{\text{rank}}(K)$ is*

$$\widehat{\text{df}}(\hat{M}^{\text{rank}}(K)) = (m_1 + m_2 - K)K + 2\sum_{k=1}^{K} \sum_{l=K+1}^{m_1} \frac{\sigma_l^2}{\sigma_k^2 - \sigma_l^2}. \tag{1}$$

Several interesting observations can be made from Theorem 1. First of all, it indicates that the number of free parameters in specifying a low rank matrix underestimates the degrees of freedom for $\hat{M}(K)$. To see this, note that the number of free parameters to specify an $m_1 \times m_2$ matrix of rank K is $(m_1 + m_2 - K)K$, i.e., the first term on the right hand side of (1). Because the second term on right hand side of (1) is always nonnegative,

$$\text{df}(\hat{M}^{\text{rank}}(K)) = \mathbb{E}\widehat{\text{df}}(\hat{M}^{\text{rank}}(K)) \geq (m_1 + m_2 - K)K.$$

Moreover, since with probability one $\sigma_{m_1} > 0$, the inequality is strict unless $K = 0$ or m_1. To further demonstrate the necessity of the bias correction, we now conduct a small numerical experiment.

In this experiment, we fix $m_1 = m_2 = 50$, and the underlying truth $M = AB^{\mathsf{T}}$ where A and B are independently sampled from $N(0, \mathbf{I}_{50} \otimes \mathbf{I}_5)$ so that M has rank five. We then

simulate $Y \sim \mathrm{N}(M, \mathbf{I}_{50} \otimes \mathbf{I}_{50})$ and compute $\hat{M}(K)$ for $K = 1, 2, \ldots, 50$. We compare three different ways of measuring the complexity of $\hat{M}(K)$:

- True degrees of freedom – $\mathbb{E}\langle \hat{M}(K), E \rangle$ with the expectation estimated from 1000 simulated datasets;
- Unbiased estimate of degrees of freedom – $\widehat{\mathrm{df}}(\hat{M}^{\mathrm{rank}}(K))$ as given by (1);
- Naive estimate of degrees of freedom – Number of free parameters needed to specify a rank K matrix.

The left panel of Figure 1 gives the degrees of freedom along with its two estimates for a typical simulated dataset. It is clear that the unbiased estimate given in Theorem 1 is much more accurate than the naive estimate. To further confirm the unbiasedness of $\widehat{\mathrm{df}}(\hat{M}^{\mathrm{rank}}(K))$. We repeat the experiment 1000 times and compute the sample expectation of both estimates. As shown in the right panel of Figure 1, $\mathbb{E}\widehat{\mathrm{df}}(\hat{M}^{\mathrm{rank}}(K))$ agrees with the true degrees of freedom fairly well.

To appreciate the practical implications of the unbiasedness of $\widehat{\mathrm{df}}(\hat{M}^{\mathrm{rank}}(K))$. We consider using the unbiased risk estimate $C_p(\hat{M}^{\mathrm{rank}}(K))$ to select the appropriate rank K. When using $\widehat{\mathrm{df}}(\hat{M}^{\mathrm{rank}}(K))$ as the estimated degrees of freedom, $K = 5$ is correctly identified for all of the 1000 runs. In contrast, when using $(m_1 + m_2 - K)K$ as the degrees of freedom, the correct rank is chosen only 85% of the time. For the remaining 15% runs, the selected rank is greater than $K = 5$. This may be attributed to the downward bias of the naive degrees of freedom estimate and agrees with our earlier findings.

2.3 Nuclear Norm Penalization and Its Degrees of Freedom

Alternatively to the rank constraint, nuclear norm regularization is also widely used for low rank matrix estimation:

$$\hat{M}^{\mathrm{nuclear}}(\lambda) = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\operatorname{argmin}} \left(\frac{1}{2} \|A - Y\|_{\mathrm{F}}^2 + \lambda \|A\|_* \right),$$

where $\lambda \geq 0$ is a tuning parameter, and $\|\cdot\|_*$ stands for the matrix nuclear norm, i.e.,

$$\|Y\|_* = \sum_{k=1}^{m_1} \sigma_k.$$

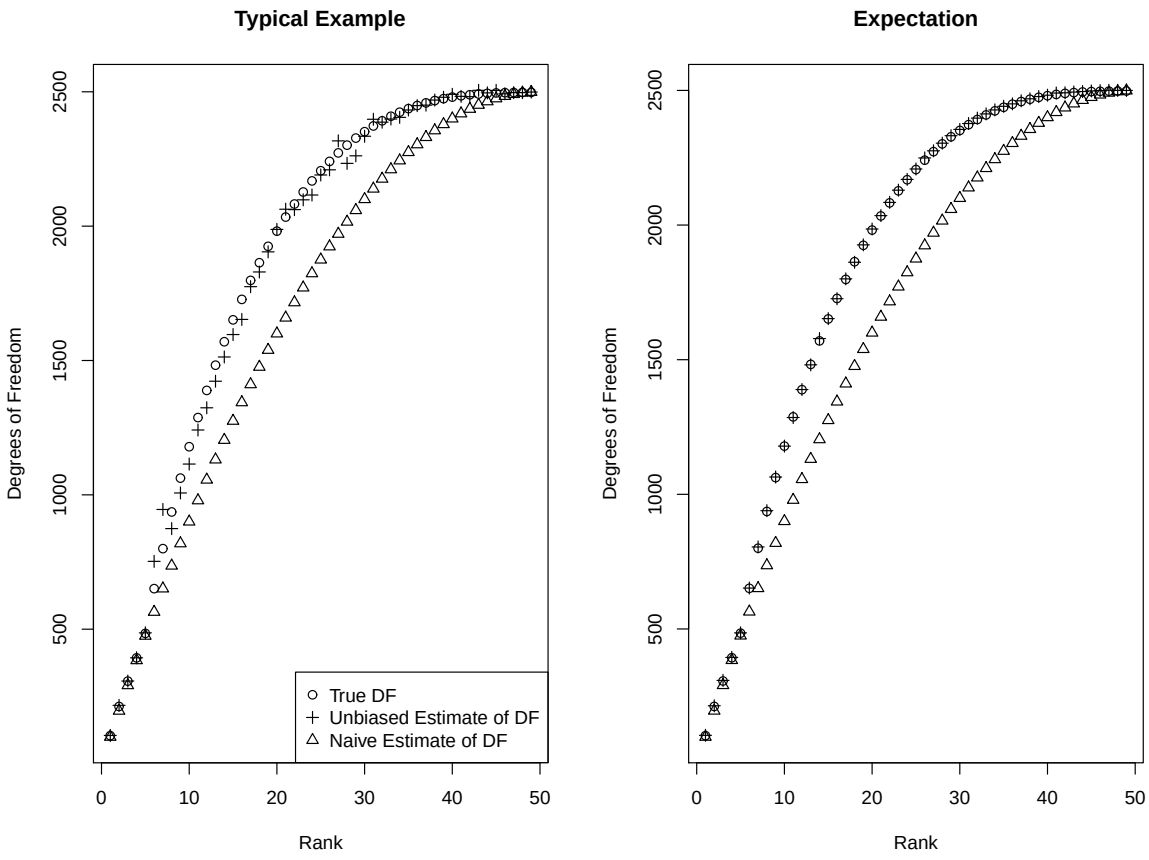

Figure 1: Degrees of freedom for reduced rank estimators: circles stand for the true degrees of freedom; pluses represent the unbiased estimate of the degrees of freedom; triangles correspond to the naive count of number of free parameters. The left panel is from a typical simulated dataset and right hand side is based on results averaged over 1000 simulations.

Similar to the rank constrained estimate, $\hat{M}^{\text{nuclear}}(\lambda)$ can be expressed in closed form:

$$\hat{M}^{\text{nuclear}}(\lambda) = \sum_{k=1}^{m_1} (\sigma_k - \lambda)_+ \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}},$$

where $(x)_+ = \max\{x, 0\}$. Nuclear norm regularization also allows for closed-form degrees of freedom estimator.

Theorem 2 *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m_1} \geq 0$ be the singular values of $Y = M + E$ where $E \sim \mathrm{N}(0, \tau^2 \mathbf{I}_{m_1} \otimes \mathbf{I}_{m_2})$ such that $\sigma_K > \lambda \geq \sigma_{K+1}$. Then an unbiased estimator of the degrees of freedom for $\hat{M}^{\text{nuclear}}(\lambda)$ is*

$$\begin{aligned} \widehat{\mathrm{df}}(\hat{M}^{\text{nuclear}}(\lambda)) &= (m_1 + m_2 - K)K + 2\sum_{k=1}^{K}\sum_{l=K+1}^{m_1} \frac{\sigma_l^2}{\sigma_k^2 - \sigma_l^2} \\ &\quad -\lambda(m_2 - m_1)\sum_{k=1}^{K}\frac{1}{\sigma_k} - 2\lambda\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_k}{\sigma_k^2 - \sigma_l^2}. \end{aligned} \tag{2}$$

Comparing (1) and (2), one recognizes that the first two terms on the right hand side of (2) correspond to the degrees of freedom for the rank constrained estimator of the same rank. The remaining two terms specify how much less complexity a nuclear norm regularized estimator has when compared with rank constrained estimator of the same rank.

We note that the number of free parameters in specifying a low rank matrix again incorrectly measures the complexity of $\hat{M}^{\text{nuclear}}(\lambda)$ because

$$\begin{aligned} \widehat{\mathrm{df}}(\hat{M}^{\text{nuclear}}(\lambda)) &= (m_1 + m_2 - K)K + 2\sum_{k=1}^{K}\sum_{l=K+1}^{m_1}\left(\frac{\sigma_l^2}{\sigma_k^2 - \sigma_l^2} - \frac{\lambda\sigma_k}{\sigma_k^2 - \sigma_l^2}\right) \\ &\quad -\lambda(m_2 - m_1)\sum_{k=1}^{K}\frac{1}{\sigma_k} - 2\lambda\sum_{\substack{k,l=1\\ l\neq k}}^{K}\frac{\sigma_k}{\sigma_k^2 - \sigma_l^2} \\ &= (m_1 + m_2 - K)K - 2\sum_{k=1}^{K}\sum_{l=K+1}^{m_1}\frac{\lambda\sigma_k - \sigma_l^2}{\sigma_k^2 - \sigma_l^2} \\ &\quad -\lambda(m_2 - m_1)\sum_{k=1}^{K}\frac{1}{\sigma_k} - 2\lambda\sum_{1\leq k<l\leq K}\frac{1}{\sigma_k + \sigma_l} \\ &\leq (m_1 + m_2 - K)K, \end{aligned}$$

where the inequality is strict with probability one when $K > 1$ and $K < m_1$.

To further illustrate this observation, we repeat the experiment from the previous subsection. This time we apply the nuclear norm penalization to each simulated dataset. In the left panel of Figure 2, we plot the true degrees of freedom, the proposed unbiased estimate and the naive estimate by counting the number of free parameters needed to specify a low rank matrix for a typical simulated dataset. It is clear that the unbiased estimator proposed here enjoys superior performance and the naive estimate overestimate the complexity of the nuclear norm penalization. The right panel of Figure 2 presents the results averaged over 1000 runs. It again shows the unbiasedness of $\widehat{\mathrm{df}}(\hat{M}^{\mathrm{nuclear}}(\lambda))$ given in (2).

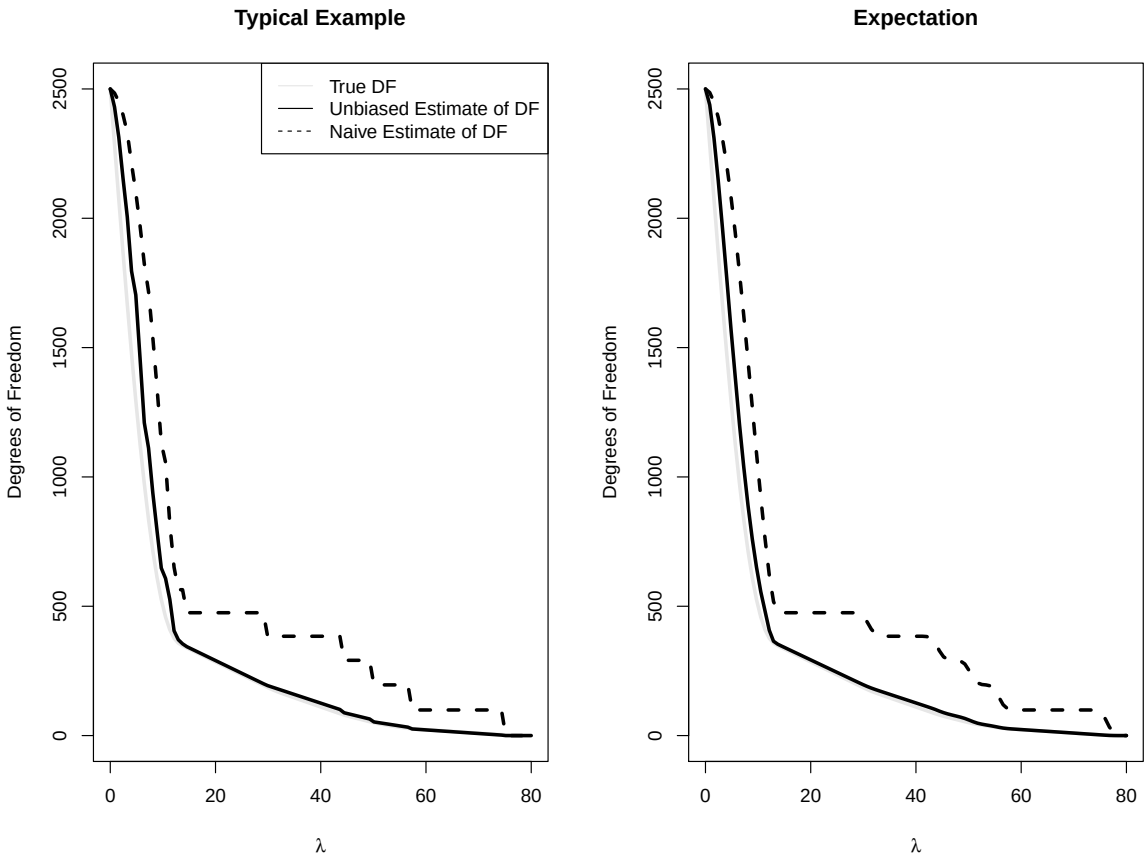

Figure 2: Degrees of freedom for nuclear norm penalization: solid grey lines correspond to the true degrees of freedom; solid black lines represent the unbiased estimate of the degrees of freedom; and the dashed block lines correspond to the naive count of number of free parameters. The left panel is from a typical simulated dataset and right hand side is based on results averaged over 1000 simulations.

The characterization of the degrees of freedom for nuclear norm penalization provided in Theorem 2 also has important practical implications. Clearly the performance of the nuclear norm penalization depends critically on the choice of the tuning parameter λ. In practice, λ is often selected by optimizing a performance evaluation or model selection criterion such as C_p or GCV. Such a criterion typically can be expressed as a bivariate function of the goodness-of-fit $\|Y - \hat{M}^{\text{nuclear}}(\lambda)\|_{\text{F}}^2$ and the degrees of freedom, i.e., $\mathcal{C}(\|Y - \hat{M}^{\text{nuclear}}(\lambda)\|_{\text{F}}^2, \widehat{\text{df}}(\hat{M}^{\text{nuclear}}(\lambda)))$ in such a way that $\mathcal{C}$ is an increasing function of both arguments. One then chooses an λ that minimizes $\mathcal{C}$. The following corollary of Theorem 2 shows that for such a purpose, it suffices to consider a finite number of choices for λ. Since $\hat{M}^{\text{nuclear}}(\lambda) = \mathbf{0}$ for all $\lambda \geq \sigma_1$, we shall assume that $0 \leq \lambda \leq \sigma_1$ without loss of generality.

Corollary 3 *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m_1}$ be the singular values of $Y = M + E$ where $E \sim \text{N}(0, \tau^2 \mathbf{I}_{m_1} \otimes \mathbf{I}_{m_2})$. Denote by*

$$\hat{\lambda} = \underset{0 \leq \lambda \leq \sigma_1}{\text{argmin}}\, \mathcal{C}(\|Y - \hat{M}^{\text{nuclear}}(\lambda)\|_{\text{F}}^2, \widehat{\text{df}}(\hat{M}^{\text{nuclear}}(\lambda))),$$

where $\mathcal{C}$ is an increasing function of both of its arguments and $\widehat{\text{df}}(\hat{M}^{\text{nuclear}}(\lambda))$ is given by (2), then

$$\hat{\lambda} \in \{\sigma_1, \sigma_2, \ldots, \sigma_{m_1}\}.$$

3 Other Low Rank Matrix Estimation Problems

Thus far, we have focused on the canonical low rank matrix estimation problem. The technique we developed, however, can be extended to other related problems as well. We now consider a couple examples.

3.1 Multivariate Linear Regression and Reduced Rank Regression

One of the most classical examples of low rank matrix estimation is the reduced rank regression for multivariate linear regression (see, e.g., Reinsel and Velu, 1998). Consider the following multivariate linear regression:

$$Y = XM + E,$$

where $Y = (y_1, \ldots, y_n)^\mathsf{T}$ is an $n \times q$ response matrix, $X = (x_1, \ldots, x_n)^\mathsf{T}$ is an $n \times p$ covariate matrix, M is a $p \times q$ coefficient matrix, and the regression noise $E \sim \mathrm{N}(0, \tau^2 \mathbf{I}_n \otimes \mathbf{I}_q)$. Let $\hat{M}$ be an estimator of M, then the fitted value can be given as $\hat{Y} = X\hat{M}$. It is clear that when $X = \mathbf{I}$, the multivariate linear regression becomes the canonical low rank matrix estimation problem investigated in the previous section. Following the same rationale as before, the prediction performance of $\hat{M}$ can be assessed using the following C_p type statistic:

$$C_p(\hat{M}) = \|Y - X\hat{M}\|_{\mathrm{F}}^2 + 2\tau^2 \mathrm{df}(\hat{M}),$$

where the degrees of freedom for $\hat{M}$ is defined as

$$\mathrm{df}(\hat{M}) := \frac{1}{\tau^2} \sum_{i=1}^{n} \sum_{j=1}^{q} \mathrm{cov}(\hat{Y}_{ij}, Y_{ij}).$$

Low rank estimation has been studied extensively in the context of multivariate linear regression. Numerous methods have been proposed over the year. See Hotelling (1935; 1936), Anderson (1951), Massy (1965), Izenman (1975), Wold (1975), Frank and Friedman, (1993), Brooks and Stone (1994), Breiman and Friedman (1997), Yuan et al. (2007) and Bunea et al. (2011) among many others. In particular, reduced rank regression is one of the most commonly used in practice (see, e.g., Reinsel and Velu, 1998). The reduced rank regression estimate of M is given by

$$\hat{M}^{\mathrm{RR}}(K) := \underset{A \in \mathbb{R}^{m_1 \times m_2} : \mathrm{rank}(A) \le K}{\mathrm{argmin}} \|Y - XA\|_{\mathrm{F}}^2.$$

The estimate $\hat{M}^{\mathrm{RR}}(K)$ can be written explicitly as

$$\hat{M}^{\mathrm{RR}}(K) = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y V V^\mathsf{T},$$

where $V = (V_1, \ldots, V_K)$ and V_k is the kth eigenvector of $Y^\mathsf{T} X (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y$.

The following theorem shows that analytic forms for the unbiased estimator of the degrees of freedom also exist in reduced rank regression.

Theorem 4 *Let $\lambda_1 \ge \lambda_2 \ge \lambda_K > \lambda_{K+1} \ge \ldots \ge \lambda_m$ be the eigenvalues of $Y^\mathsf{T} X (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y$ where $m = \min\{p, q\}$. Then an unbiased estimator of the degrees of freedom for $\hat{M}^{\mathrm{RR}}(K)$ is*

$$\widehat{\mathrm{df}}(\hat{M}^{\mathrm{RR}}(K)) = (p + q - K)K + 2 \sum_{k=1}^{K} \sum_{l=K+1}^{m} \frac{\lambda_l}{\lambda_k - \lambda_l}. \tag{3}$$

3.2 Matrix Completion and Nuclear Norm Penalization

We now turn to the problem of matrix completion under uniform sampling at random. The goal is to recover a low random matrix $M \in \mathbb{R}^{m_1 \times m_2}$ ($m_1 \leq m_2$) based on n independent random pairs (X_i, Y_i), $i = 1, 2, \ldots, n$, satisfying

$$Y_i = \langle X_i, M \rangle + \epsilon_i,$$

where the observational noise ϵ_i are i.i.d. $\mathrm{N}(0, \tau^2)$, and X_is are i.i.d. following a uniform distribution over

$$\mathcal{X} := \left\{ e_j(m_1) e_k(m_2)^{\mathsf{T}} : 1 \leq j \leq m_1, 1 \leq k \leq m_2 \right\},$$

and $e_j(m)$ is the jth canonical basis for $\mathbb{R}^m$. Problems of this type have received considerable attention in the past several years. See Candés and Recht (2008), Candés and Tao (2009), Candés and Plan (2009), Recht (2010), Gross (2011), Rohde and Tsybakov (2011), and Koltchinskii, Lounici and Tsybakov (2011) among others.

We shall consider here in particular the following version of nuclear norm penalization introduced by Koltchinskii, Lounici and Tsybakov (2011):

$$\hat{M}(\lambda) = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|A\|_{\mathrm{F}}^2 - \left\langle \frac{m_1 m_2}{n} \sum_{i=1}^{n} Y_i X_i, A \right\rangle + \lambda \|A\|_* \right\}$$

As shown by Koltchinskii et al. (2011), when λ is chosen appropriately, the resulting estimate can achieve nearly optimal rate of convergence. The practical difficulty here of course is how to select λ, which as we argued before, oftentimes relies on a good estimate of the degrees of freedom for $\hat{M}(\lambda)$. As in the multivariate linear regression setting, the degrees of freedom for the matrix completion problem can be defined as

$$\mathrm{df}(\hat{M}(\lambda)) := \frac{1}{\tau^2} \sum_{i=1}^{n} \mathrm{cov}(\hat{Y}_i, Y_i),$$

where $\hat{Y}_i = \langle X_i, \hat{M} \rangle$. The following theorem provides explicit forms of the unbiased estimate of the degrees of freedom for $\hat{M}(\lambda)$.

Theorem 5 *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{m_1}$ be the singular values of $(m_1 m_2/n) \sum_{i=1}^{n} Y_i X_i$ such that $\sigma_K > \lambda \geq \sigma_{K+1}$, and $\mathbf{u}_k$ and $\mathbf{v}_k$ be the left and right singular vectors corresponding to*

σ_k. *Then an unbiased estimator of the degrees of freedom for* $\hat{M}(\lambda)$ *is*

$$\begin{aligned}\widehat{\mathrm{df}}(\hat{M}(\lambda)) &= \frac{m_1 m_2}{n}\sum_{i=1}^{n}\mathrm{trace}\Bigg[\sum_{k:\sigma_k>\lambda}\left(1-\frac{\lambda}{\sigma_k}\right)\left(\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}X_iX_i^{\mathsf{T}}+X_i^{\mathsf{T}}X_i\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right)\\ &\quad+\sum_{k:\sigma_k>\lambda}\left(\frac{2\lambda}{\sigma_k}-1\right)X_i^{\mathsf{T}}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}X_i\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\\ &\quad+\sum_{k:\sigma_k>\lambda}\sum_{l:l\neq k}\frac{(\sigma_k-\lambda)\sigma_l}{\sigma_k^2-\sigma_l^2}\left(X_i^{\mathsf{T}}\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}X_i^{\mathsf{T}}\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}}+X_i^{\mathsf{T}}\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}}X_i^{\mathsf{T}}\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}\right)\\ &\quad+\sum_{k:\sigma_k>\lambda}\sum_{l:l\neq k}\frac{(\sigma_k-\lambda)\sigma_l^2}{\sigma_k(\sigma_k^2-\sigma_l^2)}\left(X_i^{\mathsf{T}}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}X_i\mathbf{v}_l\mathbf{v}_l^{\mathsf{T}}+X_i^{\mathsf{T}}\mathbf{u}_l\mathbf{u}_l^{\mathsf{T}}X_i\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right)\Bigg]. \qquad (4)\end{aligned}$$

4 Numerical Experiments

We now conduct some numerical experiments to illustrate the practical merits of our theoretical development. We begin with a simulation study designed to demonstrate the effect of degrees of freedom estimates on tuning parameter selection for both the rank constrained and nuclear norm regularized estimators. To fix ideas, we shall focus on the canonical model. More specifically, we first simulated the true mean matrix $M \in \mathbb{R}^{m_1\times m_2}$ $(m_1 = m_2 = 100)$ such that its left and right singular vectors are uniform over the Steifel manifold. Its singular values are independently sampled from a mixture distribution $0.9\delta(0)+0.1\mathcal{E}((\sqrt{m_1}+\sqrt{m_2})\alpha)$ with $\alpha = 0.5, 1, 1.5$ or 2, where $\delta(0)$ is a point mass at 0 and $\mathcal{E}(x)$ is the exponential distribution with mean x. The observation Y was then simulated from $\mathrm{N}(M, \mathbf{I}_{m_1}\otimes\mathbf{I}_{m_2})$. It is known that the largest singular value of a $m_1 \times m_2$ matrix of standard normals is approximately $\sqrt{m_1}+\sqrt{m_2}$. Therefore the value α determines the difficulty in estimating M with $\alpha = 2$ corresponding to the easiest situation whereas $\alpha = 0.5$ to the most difficult task.

We consider both the rank regularized and nuclear norm regularized estimators with tuning parameters, rank K for the rank regularized estimator and λ for the nuclear norm regularized estimator, selected by either C_p or GCV. For each criterion, we consider using either the proposed unbiased degrees of freedom estimator and the naive count of free parameters needed to specify a low rank matrix, giving a total of four selection methods for each estimator. We compare these selection methods in terms of their relative efficiency, that is,

$$\frac{\|\hat{M}(K)-M\|_{\mathrm{F}}^2}{\min_k\|\hat{M}(k)-M\|_{\mathrm{F}}^2}$$

for rank regularized estimator $\hat{M}(K)$, and

$$\frac{\|\hat{M}(\lambda) - M\|_{\mathrm{F}}^2}{\inf_\theta \|\hat{M}(\theta) - M\|_{\mathrm{F}}^2}$$

for nuclear norm regularized estimator $\hat{M}(\lambda)$. By definition, the relative efficiency of an estimator is no less than 1 and the closer it is to 1, the more accurate the corresponding estimate is. The results, based upon two hundred runs of simulation, are summarized in Figure 3.

It is evident from Figure 3 that when using the proposed unbiased degrees of freedom estimates, both C_p and GCV achieve nearly optimal performance in that their relative efficiency either equals to or is very close to 1, for both rank regularized and nuclear norm regularized estimators. Of course, in practice, we do not know the variance of the noise τ^2 and GCV may therefore provide a more attractive option. In comparison, when using the naive degrees of freedom, both C_p and GCV perform suboptimally, confirming the benefit of using a good degrees of freedom estimator.

We now consider an application to a previously published breast cancer study. The dataset, reported by Hu et al. (2006), was based on 146 Agilent 1Av2 microarrays. After initial filtering, normalization and necessary preprocessing, it contains log transformed gene expression measurements of 117 samples and 13,666 genes. Interested readers are referred to Hu et al. (2006) for details. Our interest here is in finding the possible low rank structure underlying the gene expression data. Such structures are common in gene expression studies and are the basis of many standard analysis methods (see, e.g., Alter et al., 2000; Raychaudhuri et al., 2000; Troyanskaya et al., 2001). To this end, we apply both the rank constrained and nuclear norm regularized estimators to the data. For each estimator, we consider using the GCV to select the tuning parameter. We chose GCV over C_p because it does not require the knowledge of the noise variance. As before, we consider using both the proposed unbiased estimators and the naive estimator for the degrees of freedom in GCV. The results are given in Figure 4.

For the rank constrained estimators, GCV with the unbiased degrees of freedom estimator chose a model with rank 21 whereas GCV with the naive degrees of freedom chose a model of rank 37, as indicated by the vertical grey lines in the left panel of Figure 4. Based on our theoretical development as well as the earlier simulation study, the former model might

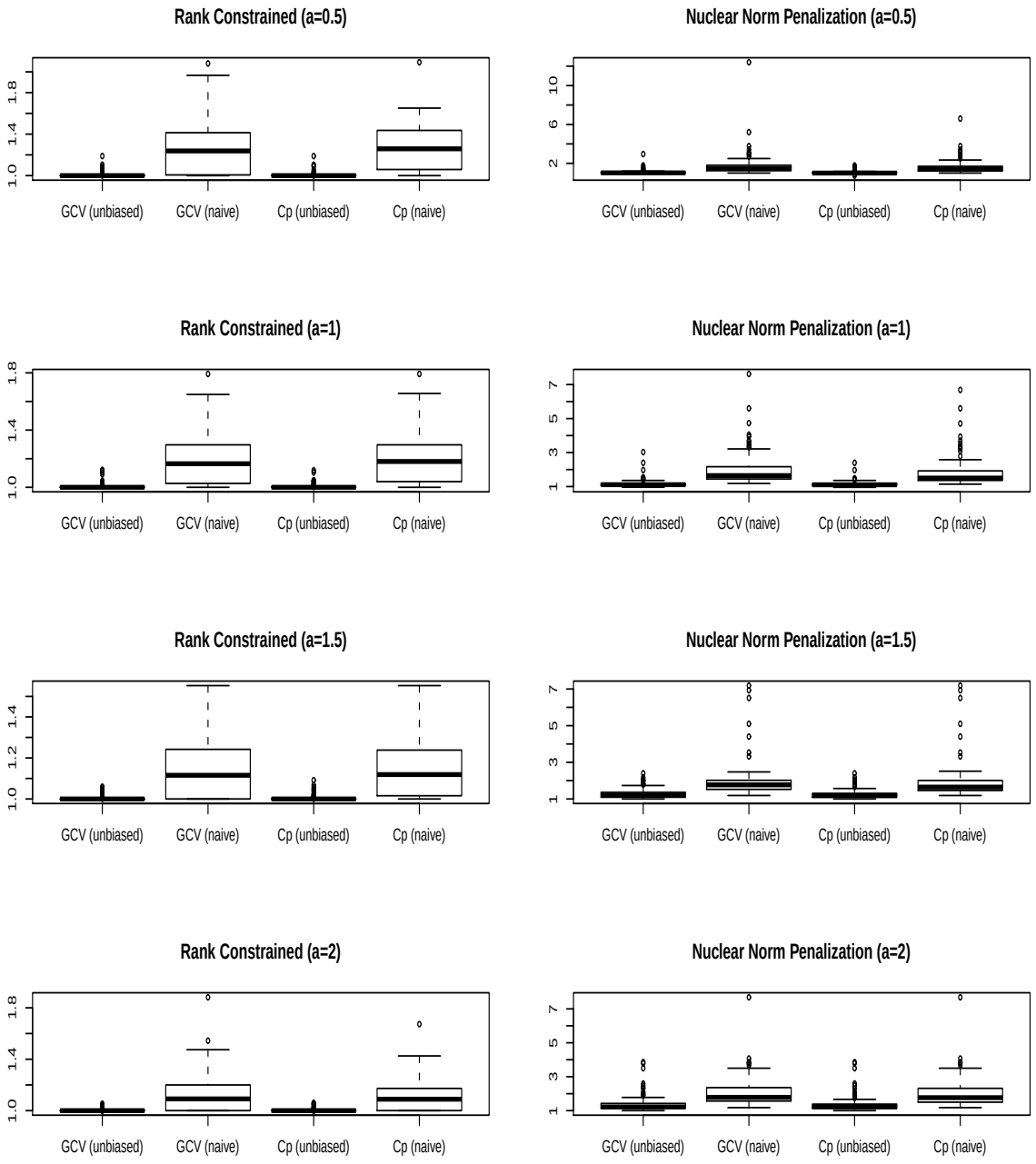

Figure 3: Effect of degrees of freedom estimators on tuning parameter selection for low rank matrix estimation: results from each panel are based on 200 simulations. For each simulated dataset, we apply rank constrained and nuclear norm regularized estimators with the tuning parameter selected by either GCV or C_p. For each tuning criterion, the degrees of freedom are estimated either by the proposed unbiased estimators or the naive count of free parameters. Reported here are the relative efficiency of each method when compared with the optimal choice of tuning parameter.

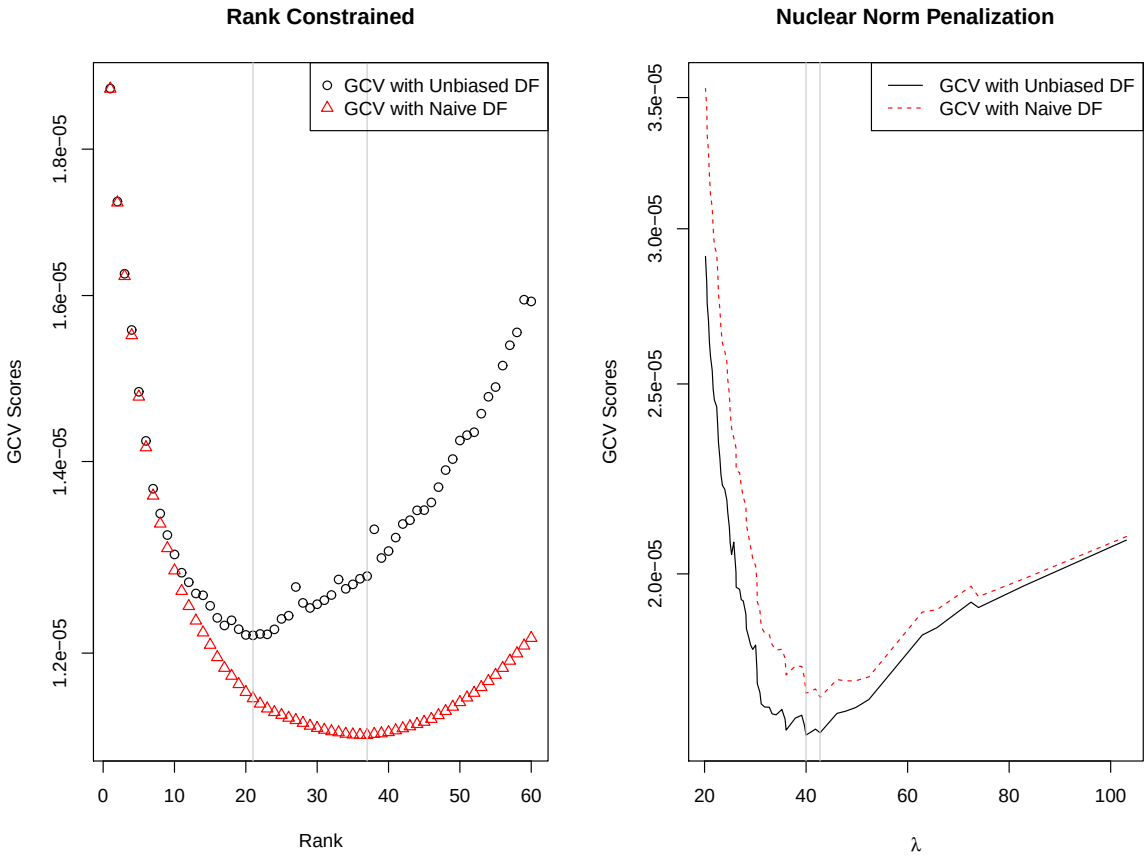

Figure 4: Breast cancer data analysis: GCV criterion was evaluated for the rank constrained and nuclear norm regularized estimators using both the proposed unbiased estimator and the naive estimator of the degrees of freedom.

be more appropriate. Similarly for the nuclear norm penalization, the tuning parameter selected with the unbiased degrees of freedom is 39.99. When using the naive degrees of freedom, the tuning parameter chosen is 42.75. Moreover, when using GCV as the model assessment criterion, rank constrained estimate is preferable for this dataset as it yields a smaller GCV score when both are appropriately tuned.

5 Proofs

5.1 Proof of Theorem 1

For brevity, we shall assume that $\tau^2 = 1$ in the proof. Write

$$J = \begin{bmatrix} 0 & Y \\ Y^\mathsf{T} & 0 \end{bmatrix},$$

the Jordan-Wielandt matrix corresponding to Y. Then $\pm\sigma_1, \pm\sigma_2, \ldots, \pm\sigma_{m_1}$ together with $m_2 - m_1$ zeros are the eigenvalues of J. Furthermore, the eigenvectors corresponding to σ_k and $-\sigma_k$ are $(\mathbf{u}_k^\mathsf{T}, \mathbf{v}_k^\mathsf{T})^\mathsf{T}/\sqrt{2} =: \eta_k$ and $(\mathbf{u}_k^\mathsf{T}, -\mathbf{v}_k^\mathsf{T})^\mathsf{T}/\sqrt{2} =: \zeta_k$ respectively. For brevity, we shall assume that there are no ties among $\sigma_1, \ldots, \sigma_{m_1}$, and $\sigma_{m_1} > 0$ in the rest of the proof. The same argument applies to the more general situation but is more tedious in notation.

By Cauchy residue formula, for a simple closed curve C in the complex plane that does not go through any of the eigenvalues of J,

$$\begin{aligned} \frac{1}{2\mathbf{i}\pi} \oint_C \frac{d\sigma}{\sigma\mathbf{I} - J} &= \sum_{k=1}^{m_1} \left(\eta_k \eta_k^\mathsf{T} \times \frac{1}{2\mathbf{i}\pi} \oint_C \frac{d\sigma}{\sigma - \sigma_k} + \zeta_k \zeta_k^\mathsf{T} \times \frac{1}{2\mathbf{i}\pi} \oint_C \frac{d\sigma}{\sigma + \sigma_k} \right) \\ &= \sum_{k:\sigma_k \in \mathrm{int}(C)} \eta_k \eta_k^\mathsf{T} + \sum_{k:-\sigma_k \in \mathrm{int}(C)} \zeta_k \zeta_k^\mathsf{T} + \left(\mathbf{I} - \sum_{k=1}^{m_1} (\eta_k \eta_k^\mathsf{T} + \zeta_k \zeta_k^\mathsf{T}) \right) \mathbb{I}\left\{0 \in \mathrm{int}(C)\right\}. \end{aligned}$$

Therefore, for any C such that its interior contains only a single eigenvalue σ_k of J,

$$\sigma_k \eta_k \eta_k^\mathsf{T} = \frac{1}{2\mathbf{i}\pi} \oint_C \frac{\sigma d\sigma}{\sigma\mathbf{I} - J},$$

again by Cauchy residue formula.

Now denote by $\tilde{J}$ the Jordan-Wielandt matrix corresponding to Y with a small perturbation – $Y + \delta A$ where $\delta > 0$ and $A \in \mathbb{R}^{m_1 + m_2}$, i.e.,

$$\tilde{J} = \begin{bmatrix} 0 & Y + \delta A \\ (Y + \delta A)^\mathsf{T} & 0 \end{bmatrix},$$

Similar to before, let $Y + \delta A = \tilde{U}\tilde{\Sigma}\tilde{V}^{\mathsf{T}}$ be its singular value decomposition. When δ is small enough, we can choose C appropriately such that $\tilde{\sigma}_k$ is also the only eigenvalue of $\tilde{J}$ that falls into the its interior. Write

$$B = \begin{bmatrix} 0 & A \\ A^{\mathsf{T}} & 0 \end{bmatrix}.$$

Following a similar argument as before, we have

$$\tilde{\sigma}_k \tilde{\eta}_k \tilde{\eta}_k^{\mathsf{T}} = \frac{1}{2\mathbf{i}\pi} \oint_C \frac{\sigma d\sigma}{\sigma \mathbf{I} - (J + \delta B)}.$$

Therefore,

$$\begin{aligned} \tilde{\sigma}_k \tilde{\eta}_k \tilde{\eta}_k^{\mathsf{T}} - \sigma_k \eta_k \eta_k^{\mathsf{T}} &= \delta \oint_C \sigma (\sigma \mathbf{I} - J)^{-1} B (\sigma \mathbf{I} - J)^{-1} d\sigma + O(\delta^2) \\ &= \delta \eta_k \eta_k^{\mathsf{T}} B \eta_k \eta_k^{\mathsf{T}} + \delta \sigma_k \eta_k \eta_k^{\mathsf{T}} B (\sigma_k \mathbf{I} - J)^{\dagger} + \delta \sigma_k (\sigma_k \mathbf{I} - J)^{\dagger} B \eta_k \eta_k^{\mathsf{T}} + O(\delta^2). \end{aligned}$$

Similarly

$$\tilde{\sigma}_k \tilde{\zeta}_k \tilde{\zeta}_k^{\mathsf{T}} - \sigma_k \zeta_k \zeta_k^{\mathsf{T}} = \delta \zeta_k \zeta_k^{\mathsf{T}} B \zeta_k \zeta_k^{\mathsf{T}} + \delta \sigma_k \zeta_k \zeta_k^{\mathsf{T}} B (\sigma_k \mathbf{I} + J)^{\dagger} + \delta \sigma_k (\sigma_k \mathbf{I} + J)^{\dagger} B \zeta_k \zeta_k^{\mathsf{T}} + O(\delta^2).$$

With slight abuse of notation, let $\hat{M}(K; A)$ be the reduced rank estimate of M with observation $M + A$. Then

$$\begin{aligned} & \frac{1}{\delta} \left(\begin{bmatrix} 0 & \hat{M}(K; E + \delta A) \\ \hat{M}(K; E + \delta A)^{\mathsf{T}} & 0 \end{bmatrix} - \begin{bmatrix} 0 & \hat{M}(K; E) \\ \hat{M}(K; E)^{\mathsf{T}} & 0 \end{bmatrix} \right) \\ = \; & \frac{1}{\delta} \sum_{k=1}^{K} \left(\tilde{\sigma}_k \tilde{\eta}_k \tilde{\eta}_k^{\mathsf{T}} - \sigma_k \eta_k \eta_k^{\mathsf{T}} \right) \\ = \; & \sum_{k=1}^{K} \left(\eta_k \eta_k^{\mathsf{T}} B \eta_k \eta_k^{\mathsf{T}} + \zeta_k \zeta_k^{\mathsf{T}} B \zeta_k \zeta_k^{\mathsf{T}} \right) \\ & \qquad + \sum_{k=1}^{K} \sigma_k \left(\eta_k \eta_k^{\mathsf{T}} B (\sigma_k \mathbf{I} - J)^{\dagger} + \zeta_k \zeta_k^{\mathsf{T}} B (\sigma_k \mathbf{I} + J)^{\dagger} \right) \\ & \qquad + \sum_{k=1}^{K} \sigma_k \left((\sigma_k \mathbf{I} - J)^{\dagger} B \eta_k \eta_k^{\mathsf{T}} + (\sigma_k \mathbf{I} + J)^{\dagger} B \zeta_k \zeta_k^{\mathsf{T}} \right) + O(\delta). \end{aligned}$$

Recall that $\eta_k = (\mathbf{u}_k^{\mathsf{T}}, \mathbf{v}_k^{\mathsf{T}})^{\mathsf{T}}/\sqrt{2}$. We get

$$\eta_k \eta_k^{\mathsf{T}} = \frac{1}{2} \begin{bmatrix} \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}} & \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}} \\ \mathbf{v}_k \mathbf{u}_k^{\mathsf{T}} & \mathbf{v}_k \mathbf{v}_k^{\mathsf{T}} \end{bmatrix}.$$

Therefore,

$$\eta_k\eta_k^{\mathsf{T}}B\eta_l\eta_l^{\mathsf{T}} = \frac{1}{4}\begin{bmatrix} * & \mathbf{u}_k\left(\mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_l + \mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_l\right)\mathbf{v}_l^{\mathsf{T}} \\ * & * \end{bmatrix}.$$

Similarly,

$$\zeta_k\zeta_k^{\mathsf{T}}B\zeta_l\zeta_l^{\mathsf{T}} = \frac{1}{4}\begin{bmatrix} * & \mathbf{u}_k\left(\mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_l + \mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_l\right)\mathbf{v}_l^{\mathsf{T}} \\ * & * \end{bmatrix},$$

and

$$\eta_k\eta_k^{\mathsf{T}}B\zeta_l\zeta_l^{\mathsf{T}} = \frac{1}{4}\begin{bmatrix} * & \mathbf{u}_k\left(\mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_l - \mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_l\right)\mathbf{v}_l^{\mathsf{T}} \\ \mathbf{v}_k\left(\mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_k - \mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_k\right)\mathbf{u}_k^{\mathsf{T}} & * \end{bmatrix}.$$

Note also that

$$\begin{aligned}(\sigma_k\mathbf{I} - J)^{\dagger} &= \sum_{l:l\neq k}\frac{\eta_l\eta_l^{\mathsf{T}}}{\sigma_k - \sigma_l} + \sum_{l=1}^{m_1}\frac{\zeta_l\zeta_l^{\mathsf{T}}}{\sigma_k + \sigma_l} + \frac{1}{\sigma_k}\left(\mathbf{I} - \sum_{l=1}^{m_1}(\eta_l\eta_l^{\mathsf{T}} + \zeta_l\zeta_l^{\mathsf{T}})\right) \\ &= \frac{1}{\sigma_k}(\mathbf{I} - \eta_k\eta_k^{\mathsf{T}}) + \sum_{l:l\neq k}\frac{\sigma_l\eta_l\eta_l^{\mathsf{T}}}{\sigma_k(\sigma_k - \sigma_l)} - \sum_{l=1}^{m_1}\frac{\sigma_l\zeta_l\zeta_l^{\mathsf{T}}}{\sigma_k(\sigma_k + \sigma_l)}.\end{aligned}$$

Similarly,

$$(\sigma_k\mathbf{I} + J)^{\dagger} = \frac{1}{\sigma_k}(\mathbf{I} - \zeta_k\zeta_k^{\mathsf{T}}) - \sum_{l=1}^{m_1}\frac{\sigma_l\eta_l\eta_l^{\mathsf{T}}}{\sigma_k(\sigma_k + \sigma_l)} + \sum_{l:l\neq k}\frac{\sigma_l\zeta_l\zeta_l^{\mathsf{T}}}{\sigma_k(\sigma_k - \sigma_l)}.$$

Thus,

$$
\begin{aligned}
& \eta_k\eta_k^{\mathsf{T}} B(\sigma_k\mathbf{I}-J)^{\dagger} + \zeta_k\zeta_k^{\mathsf{T}} B(\sigma_k\mathbf{I}+J)^{\dagger} \\
= \; & \frac{1}{\sigma_k}\left(\eta_k\eta_k^{\mathsf{T}} + \zeta_k\zeta_k^{\mathsf{T}}\right) B - \frac{1}{\sigma_k}\left(\eta_k\eta_k^{\mathsf{T}} B\eta_k\eta_k^{\mathsf{T}} + \zeta_k\zeta_k^{\mathsf{T}} B\zeta_k\zeta_k^{\mathsf{T}}\right) \\
& + \sum_{l:l\neq k} \frac{\sigma_l}{\sigma_k(\sigma_k-\sigma_l)}\left(\eta_k\eta_k^{\mathsf{T}} B\eta_l\eta_l^{\mathsf{T}} + \zeta_k\zeta_k^{\mathsf{T}} B\zeta_l\zeta_l^{\mathsf{T}}\right) \\
& - \sum_{l=1}^{m_1} \frac{\sigma_l}{\sigma_k(\sigma_k+\sigma_l)}\left(\eta_k\eta_k^{\mathsf{T}} B\zeta_l\zeta_l^{\mathsf{T}} + \zeta_k\zeta_k^{\mathsf{T}} B\eta_l\eta_l^{\mathsf{T}}\right) \\
= \; & \begin{bmatrix} * & \frac{1}{\sigma_k}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}} A \\ * & * \end{bmatrix} - \begin{bmatrix} * & \frac{1}{\sigma_k}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}} A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}} \\ * & * \end{bmatrix} \\
& + \sum_{l:l\neq k} \frac{\sigma_l}{2\sigma_k(\sigma_k-\sigma_l)} \begin{bmatrix} * & \mathbf{u}_k\left(\mathbf{v}_k^{\mathsf{T}} A^{\mathsf{T}}\mathbf{u}_l + \mathbf{u}_k^{\mathsf{T}} A\mathbf{v}_l\right)\mathbf{v}_l^{\mathsf{T}} \\ * & * \end{bmatrix} \\
& - \sum_{l=1}^{m_1} \frac{\sigma_l}{2\sigma_k(\sigma_k+\sigma_l)} \begin{bmatrix} * & \mathbf{u}_k\left(\mathbf{u}_k^{\mathsf{T}} A\mathbf{v}_l - \mathbf{v}_k^{\mathsf{T}} A^{\mathsf{T}}\mathbf{u}_l\right)\mathbf{v}_l^{\mathsf{T}} \\ * & * \end{bmatrix} \\
= \; & \begin{bmatrix} * & \frac{1}{\sigma_k}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}} A - \frac{1}{\sigma_k}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}} A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}} \\ * & * \end{bmatrix} \\
& + \sum_{l:l\neq k} \frac{\sigma_l}{\sigma_k^2-\sigma_l^2} \begin{bmatrix} * & \mathbf{u}_k\mathbf{v}_k^{\mathsf{T}} A^{\mathsf{T}}\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}} \\ * & * \end{bmatrix} + \sum_{l:l\neq k} \frac{\sigma_l^2}{\sigma_k(\sigma_k^2-\sigma_l^2)} \begin{bmatrix} * & \mathbf{u}_k\mathbf{u}_k^{\mathsf{T}} A\mathbf{v}_l\mathbf{v}_l^{\mathsf{T}} \\ * & * \end{bmatrix}
\end{aligned}
$$

Following the same argument,

$$
\begin{aligned}
& (\sigma_k\mathbf{I}-J)^{\dagger} B\eta_k\eta_k^{\mathsf{T}} + (\sigma_k\mathbf{I}+J)^{\dagger} B\zeta_k\zeta_k^{\mathsf{T}} \\
= \; & \begin{bmatrix} * & \frac{1}{\sigma_k} A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}} - \frac{1}{\sigma_k}\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}} A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}} \\ * & * \end{bmatrix} \\
& + \sum_{l:l\neq k} \frac{\sigma_l}{\sigma_k^2-\sigma_l^2} \begin{bmatrix} * & \mathbf{u}_l\mathbf{v}_l^{\mathsf{T}} A^{\mathsf{T}}\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}} \\ * & * \end{bmatrix} + \sum_{l:l\neq k} \frac{\sigma_l^2}{\sigma_k(\sigma_k^2-\sigma_l^2)} \begin{bmatrix} * & \mathbf{u}_l\mathbf{u}_l^{\mathsf{T}} A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}} \\ * & * \end{bmatrix}
\end{aligned}
$$

It can then be derived that

$$
\begin{aligned}
& \frac{1}{\delta}\left\{\hat{M}(K;E+\delta A)-\hat{M}(K;E)\right\} \\
= \; & \frac{1}{2}\sum_{k=1}^{K}\mathbf{u}_k\left(\mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_k+\mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_k\right)\mathbf{v}_k^{\mathsf{T}} \\
& +\sum_{k=1}^{K}\left(\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}A+A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}-2\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right) \\
& +\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_k\sigma_l}{\sigma_k^2-\sigma_l^2}\left(\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}}+\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}\right) \\
& +\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_l^2}{\sigma_k^2-\sigma_l^2}\left(\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_l\mathbf{v}_l^{\mathsf{T}}+\mathbf{u}_l\mathbf{u}_l^{\mathsf{T}}A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right)+O(\delta) \\
= \; & \sum_{k=1}^{K}\left(\left(\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}A+A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right)-\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right) \\
& +\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_k\sigma_l}{\sigma_k^2-\sigma_l^2}\left(\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}}+\mathbf{u}_l\mathbf{v}_l^{\mathsf{T}}A^{\mathsf{T}}\mathbf{u}_k\mathbf{v}_k^{\mathsf{T}}\right) \\
& +\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_l^2}{\sigma_k^2-\sigma_l^2}\left(\mathbf{u}_k\mathbf{u}_k^{\mathsf{T}}A\mathbf{v}_l\mathbf{v}_l^{\mathsf{T}}+\mathbf{u}_l\mathbf{u}_l^{\mathsf{T}}A\mathbf{v}_k\mathbf{v}_k^{\mathsf{T}}\right)+O(\delta).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\frac{\partial\hat{M}_{ij}}{\partial E_{ij}} &= \sum_{k=1}^{K}\left(u_{ik}^2+v_{jk}^2-u_{ik}^2v_{jk}^2\right) \\
&\quad +\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_k\sigma_l u_{ik}u_{il}v_{jk}v_{jl}}{\sigma_k^2-\sigma_l^2} \\
&\quad +\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_l^2(u_{ik}^2v_{jl}^2+u_{il}^2v_{jk}^2)}{\sigma_k^2-\sigma_l^2}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\widehat{\mathrm{df}}(\hat{M}) &= \sum_{i,j}\frac{\partial\hat{M}_{ij}}{\partial E_{ij}} \\
&= \sum_{k=1}^{K}(m_1+m_2)-K+2\sum_{k=1}^{K}\sum_{l:l\neq k}\frac{\sigma_l^2}{\sigma_k^2-\sigma_l^2} \\
&= (m_1+m_2-K)K+2\sum_{k=1}^{K}\sum_{l=K+1}^{m_1}\frac{\sigma_l^2}{\sigma_k^2-\sigma_l^2}.
\end{aligned}
$$

5.2 Proof of Theorem 2

Similar to before, it can be deducted from Cauchy residue formula that

$$\begin{aligned}
&(\tilde{\sigma}_k-\lambda)_+\tilde{\eta}_k\tilde{\eta}_k^{\mathsf{T}}-(\sigma_k-\lambda)_+\eta_k\eta_k^{\mathsf{T}}\\
=\;&\delta\mathbb{I}(\sigma_k>\lambda)\eta_k\eta_k^{\mathsf{T}}B\eta_k\eta_k^{\mathsf{T}}+\delta(\sigma_k-\lambda)_+\eta_k\eta_k^{\mathsf{T}}B(\sigma_k\mathbf{I}-J)^{\dagger}\\
&+\delta(\sigma_k-\lambda)_+(\sigma_k\mathbf{I}-J)^{\dagger}B\eta_k\eta_k^{\mathsf{T}}+O(\delta^2);
\end{aligned}$$

and

$$\begin{aligned}
&(\tilde{\sigma}_k-\lambda)_+\tilde{\zeta}_k\tilde{\zeta}_k^{\mathsf{T}}-(\sigma_k-\lambda)_+\zeta_k\zeta_k^{\mathsf{T}}\\
=\;&\delta\mathbb{I}(\sigma_k>\lambda)\zeta_k\zeta_k^{\mathsf{T}}B\zeta_k\zeta_k^{\mathsf{T}}+\delta(\sigma_k-\lambda)_+\zeta_k\zeta_k^{\mathsf{T}}B(\sigma_k\mathbf{I}+J)^{\dagger}\\
&+\delta(\sigma_k-\lambda)_+(\sigma_k\mathbf{I}+J)^{\dagger}B\zeta_k\zeta_k^{\mathsf{T}}+O(\delta^2).
\end{aligned}$$

Let $\hat{M}(\lambda;A)$ be the nuclear norm regularized estimate of M with observation $M+A$. Then

$$\begin{aligned}
&\frac{1}{\delta}\left(\begin{bmatrix}0 & \hat{M}(\lambda;E+\delta A)\\ \hat{M}(\lambda;E+\delta A)^{\mathsf{T}} & 0\end{bmatrix}-\begin{bmatrix}0 & \hat{M}(\lambda;E)\\ \hat{M}(\lambda;E)^{\mathsf{T}} & 0\end{bmatrix}\right)\\
=\;&\frac{1}{\delta}\sum_{k=1}^{m_1}\left[(\tilde{\sigma}_k-\lambda)_+\tilde{\eta}_k\tilde{\eta}_k^{\mathsf{T}}-(\sigma_k-\lambda)_+\eta_k\eta_k^{\mathsf{T}}\right]\\
=\;&\sum_{k:\sigma_k>\lambda}\left(\eta_k\eta_k^{\mathsf{T}}B\eta_k\eta_k^{\mathsf{T}}+\zeta_k\zeta_k^{\mathsf{T}}B\zeta_k\zeta_k^{\mathsf{T}}\right)\\
&\quad+\sum_{k:\sigma_k>\lambda}(\sigma_k-\lambda)\left(\eta_k\eta_k^{\mathsf{T}}B(\sigma_k\mathbf{I}-J)^{\dagger}+\zeta_k\zeta_k^{\mathsf{T}}B(\sigma_k\mathbf{I}+J)^{\dagger}\right)\\
&\quad+\sum_{k:\sigma_k>\lambda}(\sigma_k-\lambda)\left((\sigma_k\mathbf{I}-J)^{\dagger}B\eta_k\eta_k^{\mathsf{T}}+(\sigma_k\mathbf{I}+J)^{\dagger}B\zeta_k\zeta_k^{\mathsf{T}}\right)+O(\delta).
\end{aligned}$$

Following similar calculation as in the proof of Theorem 1, we get

$$\begin{aligned}
\frac{\partial\hat{M}_{ij}(E;\lambda)}{\partial E_{ij}}\;=\;&\sum_{k:\sigma_k>\lambda}\left(\frac{\sigma_k-\lambda}{\sigma_k}(u_{ik}^2+v_{jk}^2)+\left(\frac{2\lambda}{\sigma_k}-1\right)u_{ik}^2v_{jk}^2\right)\\
&+\sum_{k:\sigma_k>\lambda}\sum_{l:l\neq k}\frac{(\sigma_k-\lambda)\sigma_l u_{ik}u_{il}v_{jk}v_{jl}}{\sigma_k^2-\sigma_l^2}\\
&+\sum_{k:\sigma_k>\lambda}\sum_{l:l\neq k}\frac{(\sigma_k-\lambda)\sigma_l^2(u_{ik}^2v_{jl}^2+u_{il}^2v_{jk}^2)}{\sigma_k(\sigma_k^2-\sigma_l^2)},
\end{aligned}$$

which yields

$$\begin{aligned}\widehat{\mathrm{df}}(\lambda) &= (m_1+m_2-K)K + 2\sum_{k:\sigma_k>\lambda}\sum_{l:\sigma_l\le\lambda}\frac{\sigma_l^2}{\sigma_k^2-\sigma_l^2} \\ &\quad -\lambda(m_2-m_1)\sum_{k:\sigma_k>\lambda}\frac{1}{\sigma_k} - 2\lambda\sum_{k:\sigma_k>\lambda}\sum_{l:l\neq k}\frac{\sigma_k}{\sigma_k^2-\sigma_l^2}.\end{aligned}$$

5.3 Proof of Corollary 3

Note that for any λ such that $\sigma_k > \lambda \ge \sigma_{k+1}$,

$$\|Y - \hat{M}^{\mathrm{nuclear}}(\lambda)\|_{\mathrm{F}}^2 \ge \|Y - \hat{M}^{\mathrm{nuclear}}(\sigma_{k+1})\|_{\mathrm{F}}^2$$

and

$$\widehat{\mathrm{df}}(\hat{M}^{\mathrm{nuclear}}(\lambda)) \ge \widehat{\mathrm{df}}(\hat{M}^{\mathrm{nuclear}}(\sigma_{k+1})).$$

Because $\mathcal{C}$ is increasing in both of its argument, we get

$$\mathcal{C}(\|Y - \hat{M}^{\mathrm{nuclear}}(\lambda)\|_{\mathrm{F}}^2, \widehat{\mathrm{df}}(\hat{M}^{\mathrm{nuclear}}(\lambda))) \ge \mathcal{C}(\|Y - \hat{M}^{\mathrm{nuclear}}(\sigma_{k+1})\|_{\mathrm{F}}^2, \widehat{\mathrm{df}}(\hat{M}^{\mathrm{nuclear}}(\sigma_{k+1}))),$$

which implies the claim.

5.4 Proof of Theorem 4

Write

$$Z = (X^{\mathsf{T}}X)^{-1/2}X^{\mathsf{T}}Y.$$

and let $Z = UDV^{\mathsf{T}}$ be its singular value decomposition. Then it is easy to verify that

$$\hat{M}^{\mathrm{RR}}(K) = (X^{\mathsf{T}}X)^{-1/2}Z(K)$$

where

$$Z(K) = \sum_{k=1}^{K}\sigma_k \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}}.$$

Now recall that an unbiased estimator of the degrees of freedom is given by

$$\widehat{\mathrm{df}}(\hat{M}^{\mathrm{RR}}(K)) = \frac{1}{\tau^2}\sum_{i,j}\frac{\partial \hat{Y}_{ij}}{\partial Y_{ij}},$$

where

$$\hat{Y}(K) = X\hat{M}^{\mathrm{RR}}(K) = X(X^{\mathsf{T}}X)^{-1/2}Z(K) =: WZ(K).$$

By the chain rule of differentiation,

$$\frac{\partial \hat{Y}_{ij}}{\partial Y_{ij}} = \sum_t W_{it} \frac{\partial Z_{tj}(K)}{\partial Y_{ij}}.$$

Note that

$$Z = (X^\mathsf{T} X)^{-1/2} X^\mathsf{T} Y = W^\mathsf{T} Y.$$

Therefore,

$$\frac{\partial Z_{tj}(K)}{\partial Y_{ij}} = \sum_{s,l} \frac{\partial Z_{tj}(K)}{\partial Z_{sl}} \frac{\partial Z_{sl}}{\partial Y_{ij}} = \sum_s W_{is} \frac{\partial Z_{tj}(K)}{\partial Z_{sj}},$$

again by the chain rule. Thus,

$$\frac{\partial \hat{Y}_{ij}}{\partial Y_{ij}} = \sum_{s,t} W_{is} W_{it} \frac{\partial Z_{tj}(K)}{\partial Z_{sj}}.$$

Together with the fact that $W^\mathsf{T} W = \mathbf{I}$, we get

$$\widehat{\mathrm{df}}(\hat{M}^{\mathrm{RR}}(K)) = \frac{1}{\tau^2} \sum_{i,j,s,t} W_{is} W_{it} \frac{\partial Z_{tj}(K)}{\partial Z_{sj}} = \frac{1}{\tau^2} \sum_{j,s,t} (W^\mathsf{T} W)_{st} \frac{\partial Z_{tj}(K)}{\partial Z_{sj}} = \frac{1}{\tau^2} \sum_{j,t} \frac{\partial Z_{tj}(K)}{\partial Z_{tj}}.$$

Following the same calculation as that of Theorem 1, we can derive that

$$\widehat{\mathrm{df}}(\hat{M}^{\mathrm{KK}}(K)) = (p+q-K)K + 2 \sum_{k=1}^{K} \sum_{l=K+1}^{m} \frac{\sigma_l^2}{\sigma_k^2 - \sigma_l^2},$$

which implies the desired statement by noting that $\lambda_k = \sigma_k^2$.

5.5 Proof of Theorem 5

Write

$$Z = \frac{m_1 m_2}{n} \sum_{i=1}^{n} Y_i X_i,$$

and let $Z = UDV^\mathsf{T}$ be its singular value decomposition. Then

$$\hat{M}(\lambda) = \sum_{j=1}^{m_1} (\sigma_j - \lambda)_+ \mathbf{u}_j \mathbf{v}_j^\mathsf{T}.$$

Note that

$$\hat{Y}_i = \langle X_i, \hat{M}(\lambda) \rangle$$

Therefore,

$$\frac{\partial \hat{Y}_i}{\partial Y_i} = \left\langle X_i, \frac{\partial \hat{M}(\lambda)}{\partial Y_i} \right\rangle.$$

By the chain rule,

$$\frac{\partial \hat{M}(\lambda)}{\partial Y_i} = \sum_{s,t} \frac{\partial \hat{M}(\lambda)}{\partial Z_{st}} \frac{\partial Z_{st}}{\partial Y_i}.$$

Observe that

$$\frac{\partial Z_{st}}{\partial Y_i} = \begin{cases} m_1 m_2/n & \text{if } X_i = e_s e_t^{\mathsf{T}} \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$\frac{\partial \hat{M}_{jk}(\lambda)}{\partial Y_i} = \frac{m_1 m_2}{n} \left\langle X_i, \frac{\partial \hat{M}_{jk}(\lambda)}{\partial Z} \right\rangle.$$

Then

$$\frac{\partial \hat{Y}_i}{\partial Y_i} = \left\langle X_i, \frac{\partial \hat{M}(\lambda)}{\partial Y_i} \right\rangle = \frac{m_1 m_2}{n} \frac{\partial \langle X_i, \hat{M}(\lambda) \rangle}{\partial \langle X_i, Z \rangle}$$

In particular, if $X_i = e_s e_t^{\mathsf{T}}$, then from the proof of Theorem 1,

$$\begin{aligned}
\frac{\partial \hat{Y}_i}{\partial Y_i} &= \frac{m_1 m_2}{n} \frac{\partial \hat{M}_{st}(E;\lambda)}{\partial Z_{st}} \\
&= \frac{m_1 m_2}{n} \Bigg(\sum_{k:\sigma_k > \lambda} \left(\frac{\sigma_k - \lambda}{\sigma_k} (u_{sk}^2 + v_{tk}^2) + \left(\frac{2\lambda}{\sigma_k} - 1 \right) u_{sk}^2 v_{tk}^2 \right) \\
&\quad + \sum_{k:\sigma_k > \lambda} \sum_{l:l \neq k} \frac{(\sigma_k - \lambda)\sigma_l u_{sk} u_{sl} v_{tk} v_{tl}}{\sigma_k^2 - \sigma_l^2} \\
&\quad + \sum_{k:\sigma_k > \lambda} \sum_{l:l \neq k} \frac{(\sigma_k - \lambda)\sigma_l^2 (u_{sk}^2 v_{tl}^2 + u_{sl}^2 v_{tk}^2)}{\sigma_k (\sigma_k^2 - \sigma_l^2)} \Bigg) \\
&= \frac{m_1 m_2}{n} \mathrm{trace} \Bigg(\sum_{k:\sigma_k > \lambda} \left(1 - \frac{\lambda}{\sigma_k} \right) \left(\mathbf{u}_k \mathbf{u}_k^{\mathsf{T}} X_i X_i^{\mathsf{T}} + X_i^{\mathsf{T}} X_i \mathbf{v}_k \mathbf{v}_k^{\mathsf{T}} \right) \\
&\quad + \sum_{k:\sigma_k > \lambda} \left(\frac{2\lambda}{\sigma_k} - 1 \right) X_i^{\mathsf{T}} \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}} X_i \mathbf{v}_k \mathbf{v}_k^{\mathsf{T}} \\
&\quad + \sum_{k:\sigma_k > \lambda} \sum_{l:l \neq k} \frac{(\sigma_k - \lambda)\sigma_l}{\sigma_k^2 - \sigma_l^2} \left(X_i^{\mathsf{T}} \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}} X_i^{\mathsf{T}} \mathbf{u}_l \mathbf{v}_l^{\mathsf{T}} + X_i^{\mathsf{T}} \mathbf{u}_l \mathbf{v}_l^{\mathsf{T}} X_i^{\mathsf{T}} \mathbf{u}_k \mathbf{v}_k^{\mathsf{T}} \right) \\
&\quad + \sum_{k:\sigma_k > \lambda} \sum_{l:l \neq k} \frac{(\sigma_k - \lambda)\sigma_l^2}{\sigma_k (\sigma_k^2 - \sigma_l^2)} \left(X_i^{\mathsf{T}} \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}} X_i \mathbf{v}_l \mathbf{v}_l^{\mathsf{T}} + X_i^{\mathsf{T}} \mathbf{u}_l \mathbf{u}_l^{\mathsf{T}} X_i \mathbf{v}_k \mathbf{v}_k^{\mathsf{T}} \right) \Bigg),
\end{aligned}$$

which implies the desired claim.

References

[1] Akaike, H. (1973), Information theory and and an extension of the maximum likelihood principle, *Second International Symposium on Information Theory*, 267-281.

[2] Alter, O., Brown, P.O. and Botstein, D. (2000), Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of National Academy of Sciences of USA*, **97**, 10101-10106.

[3] Anderson, T. (1951), Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statististics*, **22**, 327-351.

[4] Breiman, L. and Friedman, J. (1997), Predicting multivariate responses in multiple linear regression, *Journal of the Royal Statistical Society, Ser. B*, **59**, 3-54.

[5] Brooks, R. and Stone, M. (1994), Joint continuum regression for multiple predictands, *Journal of the American Statistical Association*, **89**, 1374-1377.

[6] Bunea, F., She, Y. and Wegkamp, M. (2011), Optimal selection of reduced rank estimators of high-dimensional matrices, to appear in *Annals of Statistics*.

[7] Candés, E.J. and Plan, Y. (2009), Matrix completion with noise, *Proceedings of the IEEE*, **98(6)**, 925-936.

[8] Candés, E.J. and Recht, B. (2008), Exact matrix completion via convex optimization, *Foundations of Computational Mathematics*, **9**, 717-772.

[9] Candés, E.J. and Tao, T. (2009), The power of convex relaxation: Near-optimal matrix completion, *IEEE Transactions on Information Theory*, **56(5)**, 2053-2080.

[10] Craven, P. and Wahba, G. (1979), Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, **31**, 317-403.

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

[12] Efron, B. (1983), Estimation of the error rate: improvement on cross-validation, *Journal of the American Statistical Association*, **78**, 316-331.

[13] Efron, B. (1986), How biased is the apparent error rate of a prediction rule, *Journal of the American Statistical Association*, **81**, 461-470.

[14] Efron, B. (2004), The estimation of prediction error: Covariance penalty and cross-validation, *Journal of the American Statistical Association*, **99**, 619-632.

[15] Foster, D. P. and George, E. I. (1994), The risk inflation criterion for multiple regression, *Annals of Statistics*, **22**, 1947-1975.

[16] Frank, I. and Friedman, J. (1993), A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, **35**, 109-148.

[17] Gabriel, K.R. (1971), The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453-467.

[18] Gabriel, K.R. (1978) Least squares approximation of matrices by additive and multiplicative models, *Journal of the Royal Statistical Society, Ser. B*, **40**, 186-196.

[19] Gabriel, K.R. (1998), Generalised bilinear regression, *Biometrika*, **85**, 689-700.

[20] Gower, J.C. and Hand, D.J. (1996), Biplots, volume 54 of *Monographs on Statistics and Applied Probability*, London: Chapman and Hall.

[21] Gross, D. (2011), Recovering low-rank matrices from few coefficients in any basis, *IEEE Transaction on Information Theory*, **57**, 1548-1566.

[22] Harshmanm, R.A., Green, P.E., Wind, Y. and Lundy, M.E. (1982), A model for the analysis of asymmetric data in marketing research, *Marketing Science*, **1**, 205-242.

[23] Hastie, T. and Tibshirani, R. (1990), Generalized Additive Models, Chapman & Hall, London.

[24] Hoff, P.D. (2007), Model averaging and dimension selection for the singular value decomposition, *Journal of the American Statistical Association*, **102**, 674-685.

[25] Horn, R. and Johnson, C. (1991), *Topics in Matrix Analysis*, Cambridge University Press, Cambridge.

[26] Hotelling, H. (1935), The most predictable criterion, *Journal of Educational Psychology*, **26**, 139-142.

[27] Hotelling, H. (1936), Relations between two sets of variables, *Biometrika*, **28**, 321-377.

[28] Hu, Z., Fan, C., Oh, D., Marron, J., He, X., Qaqish, B., Livasy, C., Carey, L., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M., Sawyer, L., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Orrico, A., Dreher, D., Palazzo, J., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J., Ellis, M., Olopade, O., Bernard, P. and Perou, C. (2006), The molecular portraits of breast tumors are conserved across microarray platforms, *BMC Genomics*, **7**, 96.

[29] Izenman, A. (1975), Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis*, **5**, 248-264.

[30] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011), Nuclear norm penalization and optimal rates for noisy low rank matrix completion, to appear in *Annals of Statistics*.

[31] Mallows, C. (1973), Some comments on C_p, *Technometrics*, **15**, 661-675.

[32] Massey, W. (1965), Principal components regression with exporatory statistical research, *Journal of the American Statistical Association*, **60**, 234-246.

[33] Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000), Principal components analysis to summarize microarray experiments: application to sporulation time series, *Pacific Symposium of Biocomputing*, 455-466.

[34] Recht, B. (2010), A simpler approach to matrix completion, to appear in *Journal of Machine Learning Research*.

[35] Reinsel, G. and Velu, R. (1998), *Multivariate Reduced-Rank Regression*, Springer-Verlag, New York.

[36] Rohde, A. and Tsybakov, A. (2011), Estimation of high-dimensional low-rank matrices, *The Annals of Statistics*, **39**, 887-930.

[37] Shen, X., Huang, H. and Ye, J. (2004), Adaptive model selection and assessment for exponential family distributions, *Technometrics*, **46**, 306-317.

[38] Shen, X. and Ye, J. (2002), Adaptive model selection, *Journal of the American Statistical Association*, **97**, 210-221.

[39] Stein, C. (1981), Estimation of the mean of a multivariate normal distribution, *Annals of Statistics*, **9**, 1135-1151.

[40] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001), Missing value estimation methods for DNA microarrays, *Bioinformatics* **17(6)**, 520-525.

[41] Wold, H. (1975), Soft modeling by latent variables: the nonlinear iterative partial least squares approach, in *Perspectives in Probability and Statistics: Papers in Honour of M. S. Bartlett* (ed. J. Gani), Academic Press, New York.

[42] Ye, J. (1998), On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association*, **93**, 120-130.

[43] 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, Ser. B*, **69**, 329-346.

[44] Zou, H., Hastie, T. and Tibshirani, R. (2007), On the degrees of freedom of the Lasso, *The Annals of Statistics*, **35**, 2173-2192.