Published in final edited form as:
Biometrika. 2013 December 4; 100(4): 901–920. doi:10.1093/biomet/ast036.

Reduced rank regression via adaptive nuclear norm penalization

Kun Chen,
Department of Statistics, University of Connecticut, 215 Glenbrook Road, Storrs, Connecticut 06269, U.S.A

Hongbo Dong, and
Wisconsin Institutes for Discovery, University of Wisconsin, 330 N. Orchard St., Madison, Wisconsin 53715, U.S.A

Kung-Sik Chan
Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa 52242, U.S.A

Kun Chen: kun.chen@uconn.edu; Hongbo Dong: hdong6@wisc.edu; Kung-Sik Chan: kung-sik-chan@uiowa.edu

Summary

We propose an adaptive nuclear norm penalization approach for low-rank matrix approximation, and use it to develop a new reduced rank estimation method for high-dimensional multivariate regression. The adaptive nuclear norm is defined as the weighted sum of the singular values of the matrix, and it is generally non-convex under the natural restriction that the weight decreases with the singular value. However, we show that the proposed non-convex penalized regression method has a global optimal solution obtained from an adaptively soft-thresholded singular value decomposition. The method is computationally efficient, and the resulting solution path is continuous. The rank consistency of and prediction/estimation performance bounds for the estimator are established for a high-dimensional asymptotic regime. Simulation studies and an application in genetics demonstrate its efficacy.

Keywords

Low-rank approximation; Nuclear norm penalization; Reduced rank regression; Singular value decomposition

1. Introduction

Given n observations of the response $y_i \in \Re^q$ and predictor $x_i \in \Re^p$, we consider the multivariate linear regression model

$$Y = XC_0 + E, \quad (1)$$

Supplementary Material
Supplementary material available at *Biometrika* online includes additional simulation results.

where $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $X = (x_1, \ldots, x_n)^{\mathrm{T}}$, C_0 is an $p \times q$ coefficient matrix, and $E = (e_1, \ldots, e_n)^{\mathrm{T}}$ is an $n \times q$ matrix of independently and identically distributed random errors with mean zero and variance σ^2. Throughout, we write $p \wedge q = \min(p, q)$, $n \wedge q = \min(n, q)$, $r^* = r(C_0)$ and $r_{\mathrm{x}} = r(X)$, where $r(\cdot)$ denotes the rank of a matrix.

We consider the scenario in which both the predictor dimension p and response dimension q may depend on and even exceed the sample size n. Such high-dimensional regression problems are increasingly encountered. Ordinary least squares estimation is equivalent to separately regressing each response on the set of predictors, but this ignores the dependence structure of the multivariate response and may be infeasible in high-dimensional settings. The curse of dimensionality can be mitigated by assuming that the true coefficient matrix C_0 has some low-dimensional structure and employing regularization/penalization approaches for model estimation. For Gaussian data, it is appropriate to estimate C_0 by minimizing the penalized least squares criterion

$$\frac{1}{2}\mathscr{J}(C)+\mathscr{P}_\lambda(C) \quad (2)$$

with respect to $C \in \Re^{p\times q}$, where $\mathscr{J}(C)=\|Y-XC\|_F^2$ is the sum of squared errors, with $\|\cdot\|_F$ denoting the Frobenius norm, $\mathscr{P}_\lambda(\cdot)$ is some penalty function measuring the complexity of the enclosed matrix, and λ is a non-negative tuning parameter controlling the penalty.

Within this general framework, an important model is reduced rank regression (Anderson, 1951, 1999, 2002; Izenman, 1975; Reinsel & Velu, 1998), in which dimension reduction is achieved by constraining the coefficient matrix to have low rank. The classical small-p case and maximum likelihood inference for the rank-constrained approach have been extensively investigated. Recently, Bunea et al. (2011) proposed a rank selection criterion that is valid for high dimensional settings, revealing that rank-constrained estimation can be viewed as a penalized regression method (2) with a penalty proportional to the rank of C. The penalty can also be cast as an l_0 penalty in terms of the number of non-zero singular values of C, i.e., $\mathscr{P}_\lambda(C)=\lambda r(C)=\lambda\sum_{i=1}^{p\wedge q} I\{d_i(C) \neq 0\}$, where $I(\cdot)$ is the indicator function, and $d_i(\cdot)$ represents the ith largest singular value of a matrix. This results in an estimator obtained by hard-thresholded singular value decomposition; see Section 2. Yuan et al. (2007) proposed a nuclear norm penalized least squares criterion, in which the penalty is defined as $\mathscr{P}_\lambda(C)=\lambda\|C\|_*=\lambda\sum_{i=1}^{p\wedge q} d_i(C)$, where $\|\cdot\|_*$ denotes the nuclear norm. This l_1 penalty encourages sparsity among the singular values and achieves simultaneous rank reduction and shrinkage estimation (Negahban & Wainwright, 2011; Bunea et al., 2011; Lu et al., 2012). Rohde & Tsybakov (2011) investigated the theoretical properties of the Schatten-b quasi-norm penalty, which is defined as $\mathscr{P}_\lambda(C)=\lambda\sum_{i=1}^{p\wedge q} d_i^b(C)$ for $0 < b \leq 1$, and obtained non-asymptotic bounds for prediction risk. Several other extensions and theoretical developments related to reduced rank estimation exist; see, e.g., Aldrin (2000), Negahban & Wainwright (2011), Mukherjee & Zhu (2011), and Chen et al. (2012). Reduced rank methodology has connections with many popular tools including principal component

analysis and canonical correlation analysis, and has been extensively studied in matrix completion problems (Candès & Recht, 2009; Candès et al., 2011; Koltchinskii et al., 2011).

The aforementioned reduced rank approaches are closely related to the singular value decomposition (Eckart & Young, 1936; Reinsel & Velu, 1998). It is intriguing that the rank and nuclear norm penalization approaches can be viewed as l_0 and l_1 singular value penalization methods, respectively. Moreover, the squared l_2 singular value penalty is in fact a ridge penalty, because $\sum_{i=1}^{p\wedge q} d_i^2(C) = \|C\|_F^2$. Motivated by these connections and with a desire to close the gap between the l_0 and l_1 penalization schemes, we propose the adaptive nuclear norm regularization method. The adaptive nuclear norm of a matrix $C \in \Re^{p\times q}$ is defined as a weighted sum of its singular values:

$$\|C\|_{*w} = \sum_{i=1}^{p\wedge q} w_i d_i(C), \quad (3)$$

where w_is are the non-negative weights; a similar idea can be found in an unpublished 2009 University of Illinois manuscript by Jiaming Xu. We show that the adaptive nuclear norm is non-convex when the weight of the singular value decreases with the singular value, a condition needed for a meaningful regularization; see Section 2. Despite the non-convexity, we show below that the adaptive nuclear norm penalized estimator has a closed-form solution in matrix approximation problems.

Based on the proposed adaptive nuclear norm, we develop a new method for conducting simultaneous dimension reduction and coefficient estimation in high-dimensional multivariate regression. Our proposal combines two main ideas. Firstly, the proposed method builds a bridge between the l_0 and l_1 singular value penalization methods, and it can be viewed as analogous to the adaptive lasso (Tibshirani, 1996; Zou, 2006; Huang et al., 2008) developed for univariate regression. Secondly, we penalize XC rather than C, which allows the reduced rank estimation problem to be solved explicitly and efficiently. This setup was used by Klopp (2011) and Koltchinskii et al. (2011) in trace regression problems. Compared to the computationally intensive l_1 method (Yuan et al., 2007) which tends to overestimate the rank, the proposed method with the aid of some well-chosen adaptive weights may improve rank determination. Compared to the discontinuous l_0 method (Bunea et al., 2011), the proposed method results in a continuous solution path and allows a more flexible bias-variance tradeoff in model fitting.

2. Adaptive nuclear norm penalty

The adaptive nuclear norm $\|C\|_{*w}$ forms a rich class of penalty functions indexed by the weights. Clearly, it includes the nuclear norm as a special case with unit weights. Since the nuclear norm is convex and is a matrix norm, an immediate question arises as to whether or not its weighted extension (3) preserves the convexity, which is the case for the entrywise lasso and adaptive lasso penalties (Zou, 2006). However, the following theorem shows that the convexity of (3) depends on the ordering of the non-negative weights.

Theorem 1—For any matrix $C \in \Re^{p\times q}$, let $f(C) = \|C\|_{*w}$ be defined in (3). Then $f(\cdot)$ is convex in C if and only if $w_1 \geq \cdots \geq w_{p\wedge q} \geq 0$.

Hence, for the adaptive nuclear norm (3) to be a convex function, the weights must be nondecreasing with the singular value. However, for penalized estimation, the opposite is desirable, i.e., we would and shall henceforth impose the order constraint

$$0 \leq w_1 \leq \cdots \leq w_{p\wedge q}, \quad (4)$$

which ensures that a larger singular value receives a lighter penalty to help reduce the bias and a smaller singular value receives a heavier penalty to help promote sparsity. The non-convexity of $f(\cdot)$ arises from the constraint (4), so $f(\cdot)$ is no longer a matrix norm. Here is an example showing that (3) is neither convex nor concave under constraint (4). Consider $p = q = 2$, and

$$C_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad C_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Let $w_1 = 1$ and $w_2 = 2$. It can be verified that $f(C_1) = f(C_2) = f(-C_2) = 4$, while $f\{(C_1 + C_2)/2\} = 4{\cdot}5 > \{f(C_1) + f(C_2)\}/2$; also, $f\{(C_1 - C_2)/2\} = 1{\cdot}5 < \{f(C_1) + f(-C_2)\}/2$.

Consider the low-rank matrix approximation problem, $Y = C_0 + E$, a special case of model (1) where X is an identity matrix and C_0 is $n \times q$. This model provides a framework for denoising the data matrix Y. Certain low-rank estimators of C_0 can be derived from the singular value decomposition of $Y \in \Re^{n\times q}$,

$$Y = UDV^{\mathrm{T}}, \quad D = \mathrm{diag}\{d(Y)\}, \quad (5)$$

where U and V are respectively $n \times (n \wedge q)$ and $q \times (n \wedge q)$ orthonormal matrices, $d(Y) = \{d_1(Y), \ldots, d_{n\wedge q}(Y)\}^{\mathrm{T}}$ consists of the singular values of Y in descending order, and diag$(\cdot)$ denotes a diagonal matrix with the enclosed vector on its diagonal. Consider the following two kinds of estimators of C_0,

$$\mathscr{H}_\lambda(Y) = U\mathscr{H}_\lambda(D)V^{\mathrm{T}}, \quad \mathscr{H}_\lambda(D) = \mathrm{diag}\,[d_i(Y)I\{d_i(Y) > \lambda\}, i = 1, \ldots, n\wedge q], \quad (6)$$

$$\mathscr{S}_\lambda(Y) = U\mathscr{S}_\lambda(D)V^{\mathrm{T}}, \quad \mathscr{S}_\lambda(D) = \mathrm{diag}\,[\{d_i(Y) - \lambda\}_+, i = 1, \ldots, n\wedge q], \quad (7)$$

where $\lambda \geq 0$ and $x_+ = \max(0, x)$. The following proposition shows that the hard-thresholding estimator in (6) yields the l_0 rank penalized estimator (Eckart & Young, 1936), and the soft-thresholding estimator in (7) yields the l_1 nuclear norm penalized estimator (Cai et al., 2010).

Proposition 1—For any $\lambda \geq 0$ and $Y \in \Re^{n\times q}$, $\mathcal{H}_\lambda(Y)$ defined by (6) can be characterized as $\mathscr{H}_\lambda(Y)=\arg\min_C\{\|Y-C\|_F^2+\lambda^2 r(C)\}$, and $\mathcal{S}_\lambda(Y)$ in (7) can be characterized as $\mathscr{S}_\lambda(Y)=\arg\min_C\{\|Y-C\|_F^2/2+\lambda\|C\|_*\}$.

The estimator $\mathcal{H}_\lambda(Y)$ eliminates any singular values below a threshold λ, while $\mathcal{S}_\lambda(Y)$ shrinks all the singular values by λ towards zero. These operations are natural extensions of the hard/soft-thresholding rules for scalars and vectors (Donoho & Johnstone, 1995; Cai et al., 2010). In general, a soft-thresholding estimator has smaller variance but larger bias than its hard-thresholding counterpart. Soft-thresholding may, however, be preferable when data are noisy and highly correlated (Donoho & Johnstone, 1995).

The preceding discussion on the connections between different thresholding rules and penalty terms motivates us to consider the use of the adaptive nuclear norm in bridging the gap between the l_0 and l_1 singular value penalties and fine-tuning the bias-variance tradeoff. In the context of low-rank matrix approximation, we are able to obtain, with complete characterization, an explicit global minimizer of the least squares criterion penalized by a non-convex adaptive nuclear norm.

Theorem 2—For any $\lambda \geq 0$, $0 \leq w_1 \leq \cdots \leq w_{n\wedge q}$, and $Y \in \Re^{n\times q}$ with a singular value decomposition $Y = UDV^T$, a global optimal solution to the optimization problem

$$\min_C\{\frac{1}{2}\|Y-C\|_F^2+\lambda\|C\|_{*w}\} \quad (8)$$

is $\mathcal{S}_{\lambda w}(Y)$, where

$$\mathscr{S}_{\lambda w}(Y)=U\mathscr{S}_{\lambda w}(D)V^{\mathrm{T}}, \quad \mathscr{S}_{\lambda w}(D)=diag[\{d_i(Y)-\lambda w_i\}_+, i=1,\ldots,n\wedge q]. \quad (9)$$

Further, if all the nonzero singular values of Y are distinct, then $\mathcal{S}_{\lambda w}(Y)$ is the unique optimal solution.

The fact that a closed-form global minimizer can be found for the non-convex problem (8) is rather surprising. The result stems from the von Neumann's trace inequality (Mirsky, 1975); see the Appendix. Following Zou (2006), the weights can be set to be $w_i=d_i^{-\gamma}(Y)$, for i = 1, …, $n \wedge q$, where $\gamma \geq 0$ is a pre-specified constant. In this way, the order constraint (4) is automatically satisfied. We discuss a general way to construct the weights in Section 7.

3. Adaptive nuclear norm penalization in multivariate regression

3·1. Rank and nuclear norm penalized regression methods

We now consider estimating the coefficient matrix C_0, which is possibly of low rank, in the multivariate regression model (1) with an arbitrary design matrix X. Below, let $P = X(X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}$ be the projection matrix onto the column space of X and $\hat{C}_{\mathrm{L}} = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y$ the least squares estimator of C_0, where $(\cdot)^{-}$ denotes a Moore–Penrose inverse.

The results in Section 2 can be readily applied to derive certain low-rank estimator of C_0 in the general regression setting. First, consider the rank selection criterion (Bunea et al., 2011),

$$\|Y-XC\|_F^2+\lambda^2 r(C). \quad (10)$$

Minimizing (10) is the same as minimizing $L_1(C)=\|X\hat{C}_{\mathrm{L}}-XC\|_F^2+\lambda^2 r(C)$, owing to Pythagoras' theorem. We now demonstrate that this is also equivalent to minimizing $L_2(C)=\|X\hat{C}_{\mathrm{L}}-XC\|_F^2+\lambda^2 r(XC)$, which can be cast as a constrained matrix approximation problem,

$$\min_M\{\|X\hat{C}_{\mathrm{L}}-M\|_F^2+\lambda^2 r(M)\} \text{ subject to } M=XC \text{ for some } C. \quad (11)$$

However, the constrained minimizer of (11) turns out to be the same as its unconstrained counterpart. To see this, let $V\hat{D}^2\hat{V}^{\mathrm{T}}$ be the eigenvalue decomposition of $Y^{\mathrm{T}}PY = (X\hat{C}_{\mathrm{L}})^{\mathrm{T}}X\hat{C}_{\mathrm{L}}$. The singular value decomposition of $X\hat{C}_{\mathrm{L}}$ is then given by $\hat{U}D\hat{V}^{\mathrm{T}}$, where $\hat{U} = PYV\hat{D}^{-}\hat{} = X\hat{C}_{\mathrm{L}}V\hat{D}^{-}\hat{}$. It follows from Proposition 1 that the unconstrained minimizer equals

$$\mathscr{H}_\lambda(X\hat{C}_{\mathrm{L}})=\hat{U}\mathscr{H}_\lambda(\hat{D})\hat{V}^{\mathrm{T}}=X\hat{C}_{\mathrm{L}}\hat{V}\hat{D}^{-}\mathscr{H}_\lambda(\hat{D})\hat{V}^{T}=X\hat{C}_{\mathrm{H}}^{(\lambda)}, \quad (12)$$

where $\hat{C}_{\mathrm{H}}^{(\lambda)}=\hat{C}_{\mathrm{L}}\hat{V}\hat{D}^{-}\mathscr{H}_\lambda(D)\hat{V}^{\mathrm{T}}$. Therefore, $X\hat{C}_{\mathrm{H}}^{(\lambda)}$ is the desired constrained minimizer of (11). Moreover, $r\{\mathscr{H}_\lambda(\hat{D})\}=r(X\hat{C}_{\mathrm{H}}^{(\lambda)}) \le r(\hat{C}_{\mathrm{H}}^{(\lambda)}) \le r\{\mathscr{H}_\lambda(\hat{D})\}$, i.e., $r(X\hat{C}_{\mathrm{H}}^{(\lambda)})=r(\hat{C}_{\mathrm{H}}^{(\lambda)})$. Consequently, $L_1(\hat{C}_{\mathrm{H}}^{(\lambda)})=L_2(\hat{C}_{\mathrm{H}}^{(\lambda)}) \le L_2(C) \le L_1(C)$, for any $C \in \Re^{p\times q}$, i.e., $\hat{C}_{\mathrm{H}}^{(\lambda)}$ minimizes (10). It can also be shown that the set of rank-constrained estimators, obtained by minimizing $\|Y-XC\|_F^2$ subject to $r(C) \le r$, for $r = 1, \ldots, p \wedge q$, spans the solution path of (10) (Reinsel & Velu, 1998).

The nuclear norm penalized least squares criterion (Yuan et al., 2007) is

$$\frac{1}{2}\|Y-XC\|_F^2+\lambda\|C\|_*. \quad (13)$$

We denote the minimizer of (13) by $\hat{C}_{\mathrm{N}}^{(\lambda)}$, which generally does not have an explicit form. Extensive research has been devoted to this minimization problem (Cai et al., 2010; Toh & Yun, 2010). One popular algorithm is to alternate between a majorization step of the objective function and a minimization step by singular value soft-thresholding operation until convergence, but it is computationally intensive for large-scale data (Cai et al., 2010).

3·2. Adaptive nuclear norm penalized regression method

Predictive accuracy and computational efficiency are both important in high dimensional regression problems. Motivated by criteria (10) and (13), we propose to estimate C_0 by minimizing

$$\frac{1}{2}\|Y-XC\|_F^2+\lambda\|XC\|_{*w}, \quad (14)$$

where the weights $\{w_i\}$ are required to be non-negative and non-decreasing. In practice, a fore-most task of using (14) is setting proper adaptive weights. Following Zou (2006), this can be based on the least squares solution,

$$w_i=d_i^{-\gamma}(PY)=d_i^{-\gamma}(X\hat{C}_{\mathrm{L}}), \quad (15)$$

where $PY = X\hat{C}_{\mathrm{L}}$ is the projection of Y onto the column space of X and γ is a non-negative constant.

The proposed method (14) is built on two main ideas. Firstly, the criterion focuses on the fitted values XC, and encourages sparsity among the singular values of XC rather than those of C. This may yield a low-rank estimator for XC_0 and hence for C_0 (Koltchinskii et al., 2011). A prominent advantage of this setup is that the problem can then be solved explicitly and efficiently. Secondly, the adaptive penalization of the singular values allows a flexible bias-variance tradeoff: a large singular value receives a small penalty to control possible bias, and a small singular value receives a large penalty to induce sparsity and hence reduce the rank. The following corollary shows that this criterion admits an explicit minimizer.

Corollary 1—A minimizer of (14), denoted by $\hat{C}_S^{(\lambda w)}$, is given by

$$X\hat{C}_S^{(\lambda w)}=\mathscr{S}_{\lambda w}(X\hat{C}_L)=\hat{U}\mathscr{S}_{\lambda w}(\hat{D})\hat{V}^{\mathrm{T}}, \quad \hat{C}_S^{(\lambda w)}=\hat{C}_L\hat{V}\hat{D}^{-}\mathscr{S}_{\lambda w}(\hat{D})\hat{V}^{\mathrm{T}}, \quad (16)$$

where $\hat{C}_{\mathrm{L}}$ is the least squares estimator of C_0, $\hat{U}D\hat{V}^{\mathrm{T}}$ is the singular value decomposition of $X\hat{C}_{\mathrm{L}}$, and $\mathcal{S}_{\lambda w}(\cdot)$ is defined in (9).

By Pythagoras' theorem, minimizing criterion (14) is equivalent to minimizing $\|X\hat{C}_{\mathrm{L}}-XC\|_F^2/2+\lambda\|XC\|_{*w}$ with respect to C. The above result then follows directly from Theorem 2. The proposed method first projects Y onto the column space of X, i.e., $PY = X\hat{C}_{\mathrm{L}}$; the estimator is then obtained as a low-rank approximation of PY by adaptively soft-thresholding the singular values. The thresholding level is data-driven: the smaller an initially estimated singular value, the larger its thresholding level. Therefore, the estimated rank corresponds to the smallest singular value of PY that exceeds its thresholding level, i.e., $r \hat{=} \max\{r : d_r(PY) > \lambda w_r\}$. For the choice of the weights in (15), the estimated rank is

$$\hat{r} = \max\{r : d_r(PY) > \lambda^{1/(\gamma+1)}\}, \quad (17)$$

for $0 \le \lambda < d_1^{\gamma+1}(PY)$ and $r \hat{=} 0$ for $\lambda \ge d_1^{\gamma+1}(PY)$. Therefore, the plausible range of the tuning parameter is $\lambda \in [0, d_1^{\gamma+1}(PY)]$, with $\lambda = 0$ corresponding to the least squares solution and $\lambda = d_1^{\gamma+1}(PY)$ to the null solution.

The proposed estimator $\hat{C}_{\mathrm{S}}^{(\lambda w)}$ and the l_0 estimator $\hat{C}_{\mathrm{H}}^{(\lambda)}$ in (12) differ only in their estimated singular values for XC_0, but the difference can be consequential. While the solution path of the l_0 method is discontinuous and the number of possible solutions equals the maximum rank, the proposed criterion offers more flexibility in that the resulting solution path is continuous and guided by the data-driven weights. The two methods can both be efficiently computed, in contrast to the computationally intensive l_1 method in (13).

For any fixed $\lambda \ge 0$, $\hat{C}_{\mathrm{S}}^{(\lambda w)}$ can be computed by (16). To choose an optimal λ and hence an optimal solution, *K*-fold cross validation method can be used, based on predictive performance of the models (Stone, 1974). In our numerical studies, we first compute the solutions over a grid of 100 λ values equally spaced on the log scale and select the best λ value; subsequently we refine the selection process around the chosen λ value with a finer grid of 100 λ values.

4. Rank consistency and error bounds

We study the rank estimation and prediction properties of the proposed adaptive nuclear norm penalized regression method. Our theoretical analysis is built on the framework developed by Bunea et al. (2011). We mainly focus on the random weights constructed in (15), in line with the adaptive lasso method (Zou, 2006) developed for univariate regression. Similar results are obtained for any pre-specified sequence of non-random weights satisfying certain order restriction and boundedness requirements. All the proofs are given in the Appendix.

The rank of the coefficient matrix C_0, denoted as r^*, can be viewed as the number of effective linear combinations of the predictors linked to the responses. Rank determination is always a foremost task of reduced rank estimation. The quality of the rank estimator, given in (17), clearly depends on the signal to noise ratio. Following Bunea et al. (2011), we shall use the r^*th largest singular value of XC_0, i.e., $d_{r^*}(XC_0)$, to measure the signal strength, and use the largest singular value of the projected noise matrix PE, i.e., $d_1(PE)$, to measure the noise level. Intuitively, if $d_1(PE)$ is much larger than the size of the signal, some of the signal could be masked by the noise and lost during the thresholding procedure; as such, $\hat{r}$ may be much smaller than the true rank. The lemma below characterizes the limit or the true target of $\hat{r}$ and its relationship with the signal and noise levels, as well as the adaptive weights.

Lemma 1—Suppose that there exists an index $s \le r^*$ such that $d_s(XC_0) > (1 + \delta)\lambda^{1/(\gamma+1)}$ and $d_{s+1}(XC_0) \le (1 - \delta)\lambda^{1/(\gamma+1)}$ for some $\delta \in (0, 1]$. Then $\mathrm{pr}(r \hat{=} s) \ge 1 - \mathrm{pr}\{d_1(PE) \ge \delta\lambda^{1/(\gamma+1)}\}$,

where r^* is the rank of C_0, r the estimated rank given in (17), P the projection matrix onto the column space of X, E the error matrix in model (1), and γ the power parameter in the adaptive weights (15).

To achieve consistent rank estimation, we consider the following assumptions:

Assumption 1. The error matrix E has independent $N(0, \sigma^2)$ entries.

Assumption 2. For any $\theta > 0$, $\lambda = \{(1+\theta)\sigma(\sqrt{r_x} + \sqrt{q})/\delta\}^{\gamma+1}$ with δ defined in Lemma 1, and $d_r^*(XC_0) > 2\lambda^{1/(\gamma+1)}$.

Assumption 1 ensures that the noise level $d_1(PE)$ is of order $\sqrt{r_x} + \sqrt{q}$; see Lemma 2 in the Appendix (Bunea et al., 2011). Assumption 2 concerns the signal strength relative to the noise level and the appropriate rate of the tuning parameter.

Theorem 3—Suppose Assumptions 1–2 hold. Let r be the estimated rank given in (17), and $r_x = r(X)$ the rank of X. Then $\mathrm{pr}(r = r^*) \to 1$ as $r_x + q \to \infty$.

Theorem 3 shows that the proposed estimator is able to identify the correct rank with probability tending to 1 as $r_x + q$ goes to infinity. As in Bunea et al. (2011), the consistency results can be extended to the case of sub-Gaussian errors and can also be easily adapted to the case when $r_x + q$ is bounded and the sample size n goes to infinity. The rank consistency of the proposed estimator is thus valid for both classical and high-dimensional asymptotic regimes.

Our main results about the prediction performance of the proposed estimator are presented in Theorem 4. For simplicity, we write $\hat{C}_S$ for $\hat{C}_S^{(\lambda w)}$.

Theorem 4—Suppose Assumptions 1–2 hold. Let $c = d_1(XC_0)/d_r^*(XC_0) \geq 1$. Then

$$\|X\hat{C}_S - XC_0\|_F^2 \leq \frac{1+a}{1-a}\|XB - XC_0\|_F^2 + \frac{1}{a(1-a)}\{\delta\sqrt{2} + 2(2-\delta)^{-\gamma} - (2c+\delta)^{-\gamma}\}^2\lambda^{2/(\gamma+1)}r^*,$$

with probability greater than $1 - \exp\{-\theta^2(r_x + q)/2\}$, for any $0 < a < 1$ and any $p \times q$ matrix B with $r(B) \leq r^*$. Moreover, taking $B = C_0$ and $a = 1/2$ yields

$$\begin{aligned}\|X\hat{C}_S - XC_0\|_F^2 &\leq 4\{\delta\sqrt{2} + 2(2-\delta)^{-\gamma} - (2c+\delta)^{-\gamma}\}^2\lambda^{2/(\gamma+1)}r^* \\ &= 4\{\sqrt{2} + 2(2-\delta)^{-\gamma}/\delta - (2c+\delta)^{-\gamma}/\delta\}^2(1+\theta)^2\sigma^2(\sqrt{r_x} + \sqrt{q})^2 r^*\end{aligned} \quad (18)$$

with probability greater than $1 - \exp\{-\theta^2(r_x + q)/2\}$.

The established bound in (18) shows that the prediction error is bounded by $d_1^2(PE)r^*$ up to some multiplicative constant, with probability $1 - \exp\{-\theta^2(r_x + q)/2\}$, i.e., the smaller the noise level or the true rank, the smaller the prediction error. The bound is valid for any X and C_0. The estimation error bound of $\hat{C}_S$ can also be readily derived from Theorem 4, e.g.,

if $d_{r_X}(X) \geq \rho > 0$ for some constant ρ, then under Assumptions 1–2,

$$\|\hat{C}_S - C_0\|_F^2 \leq 4\rho^{-2}\{\delta\sqrt{2}+2(2-\delta)^{-\gamma}-(2c+\delta)^{-\gamma}\}^2\lambda^{2/(\gamma+1)}r^*.$$

So far, we have considered random weights based on the least squares solution, as given in (15). We now briefly outline the results for any pre-specified sequence of non-negative and nondecreasing, non-random weights $\{w_i; i = 1, \ldots, n \wedge q\}$.

Corollary 2—Suppose Assumption 1 holds, and that (i) $0 \leq w_1 \leq \cdots \leq w_{n\wedge q}$, and there exists $0 < m < \infty$ such that $w_{r^*} \leq m \leq w_{r^*+1}$, (ii) the tuning parameter $\lambda = (1 + \theta)\sigma(\sqrt{r_X} + \sqrt{q})/m$ and (iii) $d_{r^*}(XC_0) > 2\lambda m$. Then

a. $\mathrm{pr}(\hat{r} = r^*) \to 1$ *as* $r_X + q \to \infty$; *and*

b. $\|X\hat{C}_S - XC_0\|_F^2 \leq 4(\sqrt{2}+2-w_1/m)^2(1+\theta)^2(\sqrt{r_X}+\sqrt{q})^2 r^*$ *with probability greater than* $1 - \exp\{-\theta^2(r_X + q)/2\}$.

The proof is similar to that of Theorems 3 and 4 and hence is omitted.

The error bounds of the proposed estimator established in Theorem 4 and Corollary 2 are comparable to those of the l_0 rank penalized estimator and the l_1 nuclear norm penalized estimator (Bunea et al., 2011; Rohde & Tsybakov, 2011). The rate of convergence is $(r_X + q)r^*$ because $(\sqrt{r_X} + \sqrt{q})^2 \leq 2(r_X + q)$, which is the optimal minimax rate for rank sparsity under suitable regularity conditions (Rohde & Tsybakov, 2011; Bunea et al., 2012). However, the bounds for the nuclear norm penalized estimator were obtained with extra restrictions on the design matrix, and a tuning sequence for achieving the smallest mean squared error usually does not lead to correct rank recovery (Bunea et al., 2011). While both the rank selection criterion and the proposed method are able to achieve correct rank recovery and minimal mean squared error simultaneously, the latter possesses a continuous solution path produced by data-driven adaptive penalization, which may lead to improved empirical performance.

5. Robustification of reduced rank estimation

As suggested by a referee and motivated by Mukherjee & Zhu (2011), we discuss the robustification of the reduced rank methods by adding a ridge penalty. Mukherjee & Zhu (2011) proposed a reduced rank ridge regression method; see also Bunea et al. (2011) and She (2012). The shrinkage estimation induced by a ridge penalty makes the reduced rank estimation robust and hence is especially suitable when the predictors are highly correlated. The method can be viewed as minimizing the following criterion

$$\|Y - XC\|_F^2 + \lambda_1^2 r(C) + \lambda_2 \|C\|_F^2 \quad (19)$$

where $\|C\|_F^2 = \mathrm{tr}(C^{\mathrm{T}}C) = \sum_{i=1}^{p\wedge q} d_i^2(C)$, and λ_1 and λ_2 are tuning parameters. We denote the resulting robustified estimator by $\hat{C}_{\mathrm{HR}}^{(\lambda_1,\lambda_2)}$, which can be obtained by data augmentation. Specifically, letting

$$Y^* = \begin{pmatrix} Y \\ 0_{p\times q} \end{pmatrix}, \quad X^* = \begin{pmatrix} X \\ \sqrt{\lambda_2} I_{p\times p} \end{pmatrix},$$

(19) can be written as a rank selection criterion $\|Y^* - X^* C\|_F^2 + \lambda_1^2 r(C)$, whose solution is given in (12); see Mukherjee & Zhu (2011).

The adaptive nuclear norm penalization method can also be robustified by incorporating a ridge penalty term. Similar to (14), for efficient computation, we impose a ridge penalty on *XC* rather than *C*,

$$\frac{1}{2}\|Y - XC\|_F^2 + \lambda_1 \|XC\|_{*w} + \frac{\lambda_2}{2}\|XC\|_F^2. \quad (20)$$

Interestingly, criterion (20) is analogous to the adaptive elastic net criterion (Zou & Zhang, 2009) in univariate regression. We denote the minimizer of (20) by $\hat{C}_{\mathrm{SR}}^{(\lambda_1 w, \lambda_2)}$. It can be verified that

$$\hat{C}_{\mathrm{SR}}^{(\lambda_1 w, \lambda_2)} = \frac{1}{1+\lambda_2} \hat{C}_{\mathrm{S}}^{(\lambda_1 w)}, \quad (21)$$

where $\hat{C}_{\mathrm{S}}^{(\lambda_1 w)}$, defined in (16), denotes the proposed estimator in the absence of the ridge penalty.

For each fixed λ_2, solving (19) requires inverting an $p \times p$ matrix ($X^{\mathrm{T}}X + \lambda_2 I$) and performing a singular value decomposition of an $q \times q$ matrix. When p is much greater than n, the Woodbury matrix identity is useful in speeding up computation (Hager, 1989), i.e., $(X^{\mathrm{T}}X + \lambda_2 I)^{-1} = \lambda_2^{-1} I - \lambda_2^{-2} X^{\mathrm{T}} (I + \lambda_2^{-1} X X^{\mathrm{T}})^{-1} X$. Following Mukherjee & Zhu (2011), in practice we use *K*-fold cross validation to determine the optimal rank and select the optimal λ_2 from a sequence of 100 values. On the other hand, obtaining the whole solution path of (20) only requires a one-time matrix inversion and singular value decomposition. We perform a 100 × 100 grid search of (λ_1, λ_2) to obtain the final estimator.

6. Empirical studies

6·1. Simulation

We compare the prediction, estimation and rank determination of various reduced rank estimators including the nuclear norm penalized estimator $\hat{C}_{\mathrm{N}}^{(\lambda)}$ (Yuan et al., 2007), the rank penalized estimator $\hat{C}_{\mathrm{H}}^{(\lambda)}$ (Bunea et al., 2011), and our proposed adaptive nuclear norm penalized estimator $\hat{C}_{\mathrm{S}}^{(\lambda w)}$ with several choices of the weight parameter γ. The robustified versions, namely, $\hat{C}_{\mathrm{HR}}^{(\lambda_1, \lambda_2)}$ and $\hat{C}_{\mathrm{SR}}^{(\lambda_1 w, \lambda_2)}$, are also considered. For simplicity, we suppress the superscripts from the notations of the various estimators. We used the accelerated proximal

gradient algorithm implemented in Matlab by Toh & Yun (2010) for computing $\hat{C}_{\mathrm{N}}$. R code for computing $\hat{C}_{\mathrm{HR}}$ was provided by the original authors (Mukherjee & Zhu, 2011), and we modified their code to make use of the Woodbury matrix identity. We implemented all the other methods in R (R Development Core Team, 2013). All computation was done on computers with 3.4 GHz CPU, 8 GB RAM and the Linux operating system.

We consider the same simulation models as in Bunea et al. (2011). The coefficient matrix C_0 is constructed as $C_0 = bC_1C_2^{\mathrm{T}}$, where $b > 0$, $C_1 \in \Re^{p\times r^*}$, $C_2 \in \Re^{q\times r^*}$ and all entries in C_1 and C_2 are random samples from $N(0, 1)$. Two scenarios of model dimensions are considered, with $p, q < n$ and $p, q > n$. In Model I, we set $n = 100$, $p = q = 25$ and $r^* = 10$. The matrix X is constructed by generating its n rows as random samples from $N(0, \Gamma)$, where $\Gamma = (\Gamma_{ij})_{p\times p}$ and $\Gamma_{ij} = \rho^{|i-j|}$ with some $0 < \rho < 1$. In Model II, we set $n = 20$, $p = q = 25$, $r^* = 5$ and $r_{\mathrm{x}} = 10$. The matrix X is generated as $X = X_0\Gamma^{1/2}$, where Γ is defined as above, $X_0 = X_1X_2$, $X_1 \in \Re^{n\times r_{\mathrm{x}}}$, $X_2 \in \Re^{r_{\mathrm{x}}\times p}$, and all entries of X_1, X_2 are $N(0, 1)$ random samples.

The data matrix Y is then generated by $Y = XC_0 + E$, where the elements of E are $N(0, 1)$ random samples. Each simulated model is characterized by the sample size n, the number of predictors p, the number of responses q, the true model rank r^*, the rank of the design matrix r_{x}, the correlation $\rho \in \{0{\cdot}1, 0{\cdot}5, 0{\cdot}9\}$, and the signal strength $b \in \{0{\cdot}05, 0{\cdot}1, 0{\cdot}3\}$. The experiment was replicated 500 times for each parametric setting.

One way to alleviate inaccuracy in the empirical tuning parameter selection and to reveal the true potential of each method for fair comparison is to tune each method based on its predictive accuracy evaluated with a very large independently generated validation data set; this yields optimally tuned estimators. We have also tried ten-fold cross validation for selecting the tuning parameters but the results are omitted for brevity, as they are similar to or slightly worse than those of the optimal tuning procedure; see the Supplementary Material. For each method, the model accuracy is measured by the average of the scaled mean squared errors from all 500 runs, i.e., $\mathrm{Est}(\hat{C}) = 100\|C_0 - \hat{C}\|_F^2/(pq)$ for estimation, and $\mathrm{Pred}(\hat{C}) = 100\|XC_0 - X\hat{C}\|_F^2/(nq)$ for prediction. To evaluate the rank determination performance, we report the average of the estimated ranks from all runs and the percentage of correct rank identification. Tables 1 and 2 summarize the simulation results and list the average computation time per simulation run for Models I and II.

We first examine the effects of the adaptive weights on the proposed estimator $\hat{C}_{\mathrm{S}}$. For the case of equal weights, i.e., $\gamma = 0$, $\hat{C}_{\mathrm{S}}$ tends to overestimate the rank and does not have good predictive performance in most cases. The performance of $\hat{C}_{\mathrm{S}}$ is substantially better when $\gamma = 2$, which implements data-driven weights, than when $\gamma = 0$. We have also experimented with other γ values, and our results show that $\gamma = 2$ is generally a good choice; see the Supplementary Material. Henceforth we refer to the case of $\gamma = 2$ in the following comparisons.

A sharper comparison between the various estimators can be obtained by contrasting their performance on each simulated dataset. For instance, for each experimental setting, we

compare $\hat{C}_{\mathrm{H}}$ with $\hat{C}_{\mathrm{S}}$ by computing the percentage reduction in the mean squared prediction error of $\hat{C}_{\mathrm{S}}$ relative to $\hat{C}_{\mathrm{H}}$ for each of the 500 simulated datasets:

$$100 \times \frac{\mathrm{Pred}(\hat{C}_{\mathrm{H}}) - \mathrm{Pred}(\hat{C}_{\mathrm{S}})}{\mathrm{Pred}(\hat{C}_{\mathrm{H}})}\%,$$

where Pred(·) denotes the scaled mean squared prediction error of a method. Figure 1 displays the notched boxplots of the percentage reduction in the mean squared prediction error of $\hat{C}_{\mathrm{S}}$ relative to $\hat{C}_{\mathrm{H}}$ across all experimental settings, whereas Figure 2 displays those of $\hat{C}_{\mathrm{S}}$ relative to $\hat{C}_{\mathrm{N}}$. The notches in each boxplot extend $1{\cdot}58/\sqrt{500}$ times its inter-quartile range from the median (McGill et al., 1978).

The proposed estimator $\hat{C}_{\mathrm{S}}$ generally outperforms the rank penalized estimator $\hat{C}_{\mathrm{H}}$, especially in Model I. Figure 1 shows that the improvement in prediction can be substantial, when the signal is weak or moderate and the correlation among the predictors is high. For rank determination, both estimators perform well when the signal is moderate to strong and the correlation among the predictors is weak to moderate. The proposed estimator $\hat{C}_{\mathrm{S}}$, however, tends to slightly overestimate the rank.

The proposed method often outperforms nuclear norm penalized regression (Yuan et al., 2007), and is more parsimonious than the latter in both rank reduction and computation. Table 1 and Figure 2 show that in Model I, when the signal is weak and/or the correlation among the predictors is high, the nuclear norm penalized estimator $\hat{C}_{\mathrm{N}}$ performs better than $\hat{C}_{\mathrm{S}}$ in estimation and prediction. However, this gain has a price, for $\hat{C}_{\mathrm{N}}$ often overestimates the rank and is much harder to compute. In the high dimensional setting of Model II, $\hat{C}_{\mathrm{S}}$ generally enjoys similar or better predictive performance than $\hat{C}_{\mathrm{N}}$. Our findings agree with those of Bunea et al. (2011).

Table 1 shows that shrinkage estimation due to the additional ridge penalty generally enhances an estimator, especially in the presence of highly correlated predictors. However, $\hat{C}_{\mathrm{S}}$ benefits much less from the additional ridge penalty than does $\hat{C}_{\mathrm{H}}$, because, unlike the latter, $\hat{C}_{\mathrm{S}}$ is already an adaptive shrinkage estimator, owing to the soft-thresholding operation. In general, it is worthwhile to incorporate ridge penalization in order to further improve prediction, if the increased computational cost is affordable.

6·2. Application

We consider a breast cancer data set (Witten et al., 2009), consisting of gene expression measurements and comparative genomic hybridization measurements for $n = 89$ subjects. The data were used to demonstrate the effectiveness of the rank selection criterion in a preprint of Bunea et al. (2011) posted at the website http://arxiv.org. The data set is available in the R package PMA (Witten et al., 2009), and a detailed description can be found in Chin et al. (2006).

Prior studies have demonstrated that certain types of cancer are characterized by abnormal DNA copy-number changes (Pollack et al., 2002; Peng et al., 2010). It is thus of interest to

examine the relationship between DNA copy-number variations and gene expression profiles, for which multivariate regression methods can be useful. Biologically, it makes sense to regress gene expression profiles on copy-number variations because the amplification or deletion of the portion of DNA corresponding to a given gene may result in a corresponding increase or decrease in expression of that gene. The reverse approach is also meaningful, in that the resulting predictive model may identify functionally relevant copy-number variations. This approach has been shown to be promising in enhancing the limited comparative genomic hybridization data analysis with the wealth of gene expression data (Geng et al., 2011; Zhou et al., 2012). We have tried both approaches, i.e., setting 1: designating the copy-number variations of a chromosome as predictors and the gene expression profiles of the same chromosome as responses, and setting 2: reversing the roles of the predictors and the responses. We find that in setting 1, none of the methods provides an adequate fit to the data, and the rank selection criterion may even fail to pick up any signals. The reduced rank models give much better results under setting 2. We thus report only the results for setting 2.

We focus the analysis on chromosome 21, for which $p = 227$ and $q = 44$. Both the responses and predictors are standardized. We compare the various reduced rank methods by the following cross-validation procedure. The data were randomly split into a training set of size $n_{\text{train}} = 79$ and a test set of size $n_{\text{test}} = 10$. All model estimation was carried out using the training data, with the tuning parameters selected by ten-fold cross validation. We used the test data to calibrate the predictive performance of each estimator $\hat{C}$, specifically, by its mean squared prediction error $\|Y_{\text{test}} - X_{\text{test}}\hat{C}\|_F^2/(qn_{\text{test}})$, where $(Y_{\text{test}}, X_{\text{test}})$ denotes the test set. The random-splitting process was repeated 100 times to yield the average mean squared prediction error and the average rank estimate for each method; see the upper panel of Table 3.

As the number of predictors is much greater than the sample size, it is reasonable to assume that only a subset of predictors is important. Therefore, a perhaps better modeling strategy is subset multivariate regression with a selected subset of predictors. Recently, several variable selection methods have been proposed in the context of reduced rank regression (Chen et al., 2012; Chen & Huang, 2012; Bunea et al., 2012). We modified the preceding cross-validation procedure for comparing the reduced rank subset regression methods. The only modification was that for each random split, we first applied the method of Chen et al. (2012) using the training set to select a set of predictors, with which the reduced rank methods were subsequently carried out using the training set and calibrated using the test set. Since our main goal is to compare the various reduced rank methods, we omit the description of the predictor selection procedure but refer the interested reader to Chen et al. (2012) for details. The results are summarized in the lower panel of Table 3.

Table 3 shows that the proposed estimator $\hat{C}_{\text{S}}$ enjoys slightly better predictive performance than both $\hat{C}_{\text{H}}$ and $\hat{C}_{\text{N}}$. The numbers of selected predictors in the 100 splits range from 71 to 102, hence incorporating variable selection greatly reduces the number of predictors and may potentially improve model interpretation. However, in this example, reduced rank estimation using a subset of predictors results in higher mean squared prediction error than

using all predictors, uniformly for all methods, but more so for $\hat{C}_{\mathrm{H}}$ than for other methods. The nuclear norm penalized estimator $\hat{C}_{\mathrm{N}}$ generally yields a higher rank estimate than the other methods. Incorporating ridge penalization improves the predictive performance of the reduced rank methods. Particularly, $\hat{C}_{\mathrm{HR}}$ may substantially outperform its non-robust counterpart $\hat{C}_{\mathrm{H}}$. We find that both $\hat{C}_{\mathrm{N}}$ and $\hat{C}_{\mathrm{HR}}$ can be computationally intensive for large datasets, while other methods are much faster to compute. These results are consistent with the simulation findings in Section 6·1.

7. Discussion

Adaptive nuclear norm penalization can serve as a building block to study a family of singular value penalties. This is based on the connection between an adaptive l_1 penalty and many concave penalty functions (Knight & Fu, 2000; Fan & Li, 2001; Huang et al., 2008). Consider the regression problem (2) with a general singular value penalty $\mathscr{P}_\lambda(C)=\sum_{i=1}^{p\wedge q} p_\lambda\{d_i(C)\}$, where $p_\lambda(\cdot)$ is a penalty function, e.g., $p_\lambda\{d_i(C)\}=\lambda d_i^b(C)$ for some $0 < b \leq 1$ (Huang et al., 2008; Rohde & Tsybakov, 2011). In this setup the optimization of (2) can be challenging. A promising approach is to adopt a local linear approximation (Zou & Li, 2008), $p_\lambda\{d_i(C)\} \approx p_\lambda(\tilde{d}_i)+p'_\lambda(\tilde{d}_i)\{d_i(C)-\tilde{d}_i\}$, for $d_i(C) \approx \tilde{d}_i$, where $\tilde{d}_i$ is some initial estimator of $d_i(C)$. It can be seen that for fixed $\tilde{d}_i$, up to a constant, the first-order approximation yields exactly an adaptive nuclear norm penalty. This suggests that these problems may be solved by an iteratively reweighted adaptive nuclear norm penalization approach.

Incorporating an extra ridge penalty can improve reduced rank estimation (Mukherjee & Zhu, 2011; She, 2012). When combined with the adaptive nuclear norm penalty, such a criterion bears resemblance to the adaptive elastic-net criterion (Zou & Hastie, 2005; Zou & Zhang, 2009) in univariate regression. It would be interesting to investigate the theoretical properties of this approach and compare it with the nonlinear fusion of nuclear norm and ridge penalties in Owen (2007) and She (2012). Another pressing problem is to extend regularized reduced rank regression methods to generalized linear and nonparametric regression models (Yee & Hastie, 2003; Li & Chan, 2007; She, 2012). On the optimization aspect, it is interesting to study the use of adaptive nuclear norm in some classical sparse optimization areas, such as matrix completion (Candès et al., 2011).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Howell Tong, and are grateful to an associate editor and two referees for constructive comments that helped improve the paper significantly. The work was partially supported by the U.S. National Institutes of Health and National Science Foundation.

References

Aldrin M. Multivariate prediction using softly shrunk reduced-rank regression. The American Statistician. 2000; 54:29–34.

Anderson TW. Estimating linear restrictions on regression coefficients for multivariate normal distributions. Annals of Mathematical Statistics. 1951; 22:327–351.

Anderson TW. Asymptotic distribution of the reduced rank regression estimator under general conditions. Annals of Statistics. 1999; 27:1141–1154.

Anderson TW. Specification and misspecification in reduced rank regression. Sankhyâ. 2002; 64:193–205.

Bunea F, She Y, Wegkamp M. Optimal selection of reduced rank estimators of high-dimensional matrices. Annals of Statistics. 2011; 39:1282–1309.

Bunea F, She Y, Wegkamp M. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. Annals of Statistics. 2012; 40:2359–2388.

Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization. 2010; 20:1956–1982.

Candès EJ, Li X, Ma Y, Wright J. Robust principal component analysis? Journal of the ACM. 2011; 58:1–37.

Candès EJ, Recht B. Exact matrix completion via convex optimization. Found Comput Math. 2009; 9:717–772.

Chen K, Chan KS, Stenseth NC. Reduced rank stochastic regression with a sparse singular value decomposition. Journal of the Royal Statistical Society Series B. 2012; 74:203–221.

Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. Journal of the American Statistical Association. 2012; 107:1533–1545.

Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell. 2006; 10:529–541. [PubMed: 17157792]

Donoho DL, Johnstone IM. Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association. 1995; 90:1200–1224.

Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika. 1936; 1:211–218.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Franklin, J. Matrix Theory. Toronto: Dover Publications; 2000.

Geng H, Iqbal J, Chan WC, Ali HH. Virtual CGH: an integrative approach to predict genetic abnormalities from gene expression microarray data applied in lymphoma. BMC Medical Genomics. 2011; 4:32. [PubMed: 21486456]

Hager WW. Updating the inverse of a matrix. SIAM Review. 1989; 31:221–239.

Hardy, GH.; Littlewood, JE.; Pólya, G. Inequalities. Cambridge University Press; 1967.

Horn, RA.; Johnson, CR. Matrix Analysis. Cambridge University Press; 1985.

Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Annals of Statistics. 2008; 36:587–613.

Izenman AJ. Reduced-rank regression for the multivariate linear model. Journal of Multivariate Analysis. 1975; 5:248–264.

Klopp O. Rank penalized estimators for high-dimensional matrices. Electron J Statist. 2011; 5:1161–1183.

Knight K, Fu W. Asymptotics for lasso-type estimators. Annals of Statistics. 2000; 28:1356–1378.

Koltchinskii V, Lounici K, Tsybakov A. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. Annals of Statistics. 2011; 39:2302–2329.

Li MC, Chan KS. Multivaraite reduced-rank nonlinear time series modeling. Statistica Sinica. 2007; 17:139–159.

Lu Z, Monteiro RDC, Yuan M. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Math Program. 2012; 131:163–194.

McGill R, Tukey JW, Larsen WA. Variations of box plots. The American Statistician. 1978; 32:12–16.

Mirsky L. A trace inequality of John von Neumann. Monatschefte fur Mathematik. 1975; 79:303–306.

Mukherjee A, Zhu J. Reduced rank ridge regression and its kernel extensions. Statistical Analysis and Data Mining. 2011; 4:612–622. [PubMed: 22993641]

Negahban S, Wainwright MJ. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Annals of Statistics. 2011; 39:1069–1097.

Owen AB. A robust hybrid of lasso and ridge regression. Contemporary Mathematics. 2007; 443:59–71.

Peng J, Zhu J, Bergamaschi A, Han W, Noh D-Y, Pollack JR, Wang P. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann Appl Stat. 2010; 4:53–77. [PubMed: 24489618]

Pollack JR, Sørlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani RJ, Botstein D, Børresen-Dale ALL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:12963–12968. [PubMed: 12297621]

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2013.

Reinsel, GC.; Velu, P. Multivariate Reduced-rank Regression: Theory and Applications. New York: Springer; 1998.

Rohde A, Tsybakov A. Estimation of High-Dimensional Low-rank Matrices. Annals of Statistics. 2011; 39:887–930.

She Y. Reduced Rank Vector Generalized Linear Models for Feature Extraction. Statistics and Its Interface. 2012 To appear.

Stone M. Cross-validation and multinomial prediction. Biometrika. 1974; 61:509–515.

Tibshirani RJ. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B. 1996; 58:267–288.

Toh K-C, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Pacific J Optim. 2010; 6:615–640.

von Neumann J. Some matrix inequalities and metrization of matric-space. Tomsk University Review. 1937; 1:286–300.

Witten DM, Tibshirani RJ, Hastie TJ. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009; 10:515–534. [PubMed: 19377034]

Yee T, Hastie TJ. Reduced rank vector generalized linear models. Statistical Modeling. 2003:367–378.

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

Zhou Y, Zhang Q, Stephens O, Heuck CJ, Tian E, Sawyer JR, Cartron-Mizeracki MA, Qu P, Keller J, Epstein J, Barlogie B, Shaughnessy JD. Prediction of cytogenetic abnormalities with gene expression profiles. Blood. 2012; 119:148–150.

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

Zou H, Hastie TJ. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B. 2005; 67:301–320.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics. 2008; 36:1509–1533. [PubMed: 19823597]

Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics. 2009; 37:1733–1751. [PubMed: 20445770]

Appendix. Technical details

Proof of Theorem 1

First we show by a counter-example that if we have an index k such that $w_k < w_{k+1}$, then $f(\cdot)$ is non-convex. Let C and Z be diagonal $p \times p$ matrices such that $c_{ii} = i$, for $i = 1, \ldots, p$, while Z equals C but with entries switched at positions $p - k + 1$ and $p - k$ for some $1 \le k \le p - 1$ on the diagonal. It is then easy to verify that

$$f(C)=f(Z)=\sum_{i=1}^{p} w_i(p-i+1),$$
$$f\left(\frac{C+Z}{2}\right)-\frac{f(C)}{2}-\frac{f(Z)}{2}=\left(p-k+\tfrac{1}{2}\right)(w_k+w_{k+1})-(p-k+1)w_k-(p-k)w_{k+1}$$
$$=\tfrac{1}{2}(w_k+w_{k+1})-w_k>0,$$

, where $f(\cdot)$ is defined in (3). Therefore $f(\cdot)$ is non-convex.

Next we prove that $f(C) = \|C\|_{w*}$ is a convex function of $C \in \Re^{p\times q}$ for $w_1 \ge \cdots \ge w_{p\wedge q} \ge 0$. Without loss of generality, assume $p \le q$ so that we can simply write $p = p \wedge q$. First consider the case that $w_p > 0$, and define the following function on $\Re^p$:

$$w(x)=\sum_{i=1}^{p} w_i|x|_{\delta(i)}, \quad \text{(A1)}$$

where δ is a permutation of $\{1,\ldots, p\}$ determined by x such that $|x|_{\delta(1)} \ge \cdots \ge |x|_{\delta(p)}$. We claim that $w(\cdot)$ in (A1) is a symmetric gauge function (Horn & Johnson, 1985, Definition 7.4.23), i.e., it satisfies the following six conditions: (a) $w(x) \ge 0$, for any $x \in \Re$; (b) $w(x) = 0$ if and only if $x = 0$; (c) $w(ax) = |a|w(x)$, for any $a \in \Re$; (d) $w(x + y) \le w(x) + w(y)$; (e) $w(x) = w(|x|)$; (f) $w(x) = w\{\tau(x)\}$ for any τ that is a permutation of indices $\{1,\ldots, p\}$.

All conditions except (d) are trivial to verify. To prove (d), let δ, σ and τ be permutations such that $|x + y|_{\delta(i)}$, $|x|_{\sigma(i)}$ and $|y|_{\tau(i)}$ are placed in non-increasing order respectively.

$$w(x+y)=\sum_{i=1}^{p} w_i|x+y|_{\delta(i)} \le \sum_{i=1}^{p}\left(w_i|x|_{\delta(i)}+w_i|y|_{\delta(i)}\right)$$
$$\le \sum_{i=1}^{p}\left(w_i|x|_{\sigma(i)}+w_i|y|_{\tau(i)}\right)=w(x)+w(y).$$

where the second inequality is due to the Hardy–Littlewood–Pólya inequality (Hardy et al., 1967). By a straightforward application of Horn & Johnson (1985, Theorem 7.4.24), since $\|C\|_{w*} = w\{d(C)\}$ where $d(C) = \{d_1(C),\ldots, d_p(C)\}^{\mathrm{T}}$, the function $f(\cdot) = \|\cdot\|_{w*}$ defines a matrix norm and hence is convex.

For the case that $w_p = 0$, let s be the largest index such that $w_s > 0$. For $0 < \varepsilon < w_s$, consider the perturbed w that $\tilde{w}_i = w_i$, for $i = 1,\ldots, s$, and $\tilde{w}_i = \varepsilon$, for $i = s + 1,\ldots, p$. Then for any $C, Z \in \Re^{p\times q}$, $\|C + Z\|_{w*}/2 \le \|C\|_{w*}/2 + \|Z\|_{w*}/2$. By taking $\varepsilon \to 0$, $\|C + Z\|_{w*}/2 \le \|C\|_{w*}/2 + \|Z\|_{w*}/2$. Therefore $\|\cdot\|_{w*}$ is convex.

Proof of Theorem 2

We first prove that $\mathcal{S}_{\lambda w}(Y)$ is indeed a global optimal solution to (8). Below, we write h for $n \wedge q$. Let $g=\{g_i\}_{i=1}^h=d(C)$, which implies the entries of g are in non-increasing order. Since the penalty term only depends on the singular values of C, (8) can be equivalently written as:

$$\min_{g:g_1\geq\cdots\geq g_h\geq 0}\left\{\min_{C\in\Re^{n\times q},d(C)=g}\left(\frac{1}{2}\|Y-C\|_F^2+\lambda\sum_{i=1}^h w_i g_i\right)\right\}.$$

For the inner minimization, we have the inequality

$$\begin{aligned}\|Y-C\|_F^2&=\mathrm{tr}(YY^{\mathrm{T}})-2\mathrm{tr}(YC^{\mathrm{T}})+\mathrm{tr}(CC^{\mathrm{T}})\\&=\sum_{i=1}^h d_i^2(Y)-2\mathrm{tr}(YC^{\mathrm{T}})+\sum_{i=1}^h g_i^2\\&\geq\sum_{i=1}^h d_i^2(Y)-2d(Y)^{\mathrm{T}}g+\sum_{i=1}^h g_i^2.\end{aligned}$$

The last inequality is due to von Neumann's trace inequality (von Neumann, 1937; Mirsky, 1975). Equality holds when C admits the singular value decomposition $C = U\mathrm{diag}(g)V^{\mathrm{T}}$, where U and V are defined in (5) as the left and right singular matrices of Y. Then the optimization reduces to

$$\min_{g:g_1\geq\cdots\geq g_h\geq 0}\left(\sum_{i=1}^h\left[\frac{1}{2}g_i^2-\{d_i(Y)-\lambda w_i\}g_i+\frac{1}{2}d_i^2(Y)\right]\right). \quad \text{(A2)}$$

The objective function is completely separable and is minimized only when $g_i = \{d_i(Y) - \lambda w_i\}_+$. This is a feasible solution because $\{d_i(Y)\}$ is in non-increasing order, while $\{w_i\}$ is in non-decreasing order. Therefore $\mathcal{S}_{\lambda w}(Y) = U\mathrm{diag}[\{d(Y) - \lambda w\}_+]V^{\mathrm{T}}$ is a global optimal solution to (8). The uniqueness follows by the equality condition for the von Neumann's trace inequality when Y has distinct nonzero singular values, and the uniqueness of the strictly convex optimization (A2). This concludes the proof.

Proof of Lemma 1

By (17), $r \hat{>} s$ holds if and only if $d_{s+1}(PY) > \lambda^{1/(\gamma+1)}$ and $r \hat{<} s$ holds if and only if $d_s(PY) \leq \lambda^{1/(\gamma+1)}$. Then

$$\mathrm{pr}(\hat{r}\neq s)=\mathrm{pr}\left\{d_{s+1}(PY)>\lambda^{1/(\gamma+1)}\text{ or }d_s(PY)\leq\lambda^{1/(\gamma+1)}\right\}.$$

Based on Weyl's inequalities on singular values (Franklin, 2000) and observing that $PY = XC_0 + PE$, we have $d_1(PE) \geq d_{s+1}(PY) - d_{s+1}(XC_0)$ and $d_1(PE) \geq d_s(XC_0) - d_s(PY)$. Hence

$d_{s+1}(PY) > \lambda^{1/(\gamma+1)}$ implies $d_1(PE) \geq \lambda^{1/(\gamma+1)} - d_{s+1}(XC_0)$, and $d_s(PY) \leq \lambda^{1/(\gamma+1)}$ implies $d_1(PE) \geq d_s(XC_0) - \lambda^{1/(\gamma+1)}$. It then follows that

$$\mathrm{pr}(\hat{r} \neq s) \leq \mathrm{pr}\left[d_1(PE) \geq \min\{\lambda^{1/(\gamma+1)} - d_{s+1}(XC_0), d_s(XC_0) - \lambda^{1/(\gamma+1)}\}\right].$$

Note that $\min\{\lambda^{1/(\gamma+1)} - d_{s+1}(XC_0), d_s(XC_0) - \lambda^{1/(\gamma+1)}\} \geq \delta\lambda^{1/(\gamma+1)}$. This completes the proof.

Lemma 2 (Bunea et al., 2011)

Let r_X denote the rank of X and suppose Assumption 1 holds. Then for any $t > 0$, $E\{d_1(PE)\} \leq \sigma(\sqrt{r_X} + \sqrt{q})$, and $\mathrm{pr}[d_1(PE) \geq E\{d_1(PE)\} + \sigma t] \leq \exp(-t^2/2)$.

Proof of Theorem 3

When $d_{r^*}(XC_0) > 2\lambda^{1/(\gamma+1)}$, we have $d_{r^*}(XC_0) > 2\lambda^{1/(\gamma+1)} \geq (1 + \delta)\lambda^{1/(\gamma+1)}$ and $d_{r^*+1}(XC_0) = 0 \leq (1 - \delta)\lambda^{1/(\gamma+1)}$, for some $0 < \delta \leq 1$. The effective rank s defined in Lemma 1 equals the true rank, i.e., $s = r^*$, and $\min\{\lambda^{1/(\gamma+1)} - d_{r^*+1}(XC_0), d_{r^*}(XC_0) - \lambda^{1/(\gamma+1)}\} \geq \delta\lambda^{1/(\gamma+1)}$. It then follows from Lemma 2 that

$$\begin{aligned}
&\mathrm{pr}(\hat{r}=r^*) \geq 1-\mathrm{pr}\{d_1(PE) \geq \delta\lambda^{1/(\gamma+1)}\} \\
&=1-\mathrm{pr}\{d_1(PE) \geq (1+\theta)\sigma(\sqrt{r_x}+\sqrt{q})\} \\
&\geq 1-\exp\{\theta^2(r_x+q)/2\} \\
&\to 1
\end{aligned}$$

as $r_X + q \to \infty$. This completes the proof.

Proof of Theorem 4

We write h for $n \wedge q$. By the definition of $\hat{C}_S$ in (16),

$$\|Y - X\hat{C}_S\|_F^2 + 2\lambda\sum_{i=1}^{h} w_i d_i(X\hat{C}_S) \leq \|Y - XB\|_F^2 + 2\lambda\sum_{i=1}^{h} w_i d_i(XB),$$

for any $p \times q$ matrix B. Note that

$$\begin{aligned}
\|Y - X\hat{C}_S\|_F^2 &= \|Y - XC_0\|_F^2 + \|X\hat{C}_S - XC_0\|_F^2 + 2\langle E, XC_0 - X\hat{C}_S\rangle_F, \\
\|Y - XB\|_F^2 &= \|Y - XC_0\|_F^2 + \|XB - XC_0\|_F^2 + 2\langle E, XC_0 - XB\rangle_F.
\end{aligned}$$

Then we have

$$\begin{aligned}\|X\hat{C}_{\mathrm{S}}-XC_0\|_F^2 &\le \|XB-XC_0\|_F^2+2<E,X\hat{C}_{\mathrm{S}}-XB>_F+2\lambda\{\sum_{i=1}^{h}w_i d_i(XB)-\sum_{i=1}^{h}w_i d_i(X\hat{C}_{\mathrm{S}})\}\\ &\le \|XB-XC_0\|_F^2+2<PE,X\hat{C}_{\mathrm{S}}-XB>_F+2\lambda\{\sum_{i=1}^{h}w_i d_i(XB)-\sum_{i=1}^{h}w_i d_i(X\hat{C}_{\mathrm{S}})\}\\ &\le \|XB-XC_0\|_F^2+2d_1(PE)\|X\hat{C}_{\mathrm{S}}-XB\|_*+2\lambda\{\sum_{i=1}^{h}w_i d_i(XB)-\sum_{i=1}^{h}w_i d_i(X\hat{C}_{\mathrm{S}})\}\\ &\le \|XB-XC_0\|_F^2+2d_1(PE)_r(X\hat{C}_{\mathrm{S}}-XB)^{1/2}\|X\hat{C}_{\mathrm{S}}-XB\|_F+2\lambda\{\sum_{i=1}^{h}w_i d_i(XB)-\sum_{i=1}^{h}w_i d_i(X\hat{C}_{\mathrm{S}})\}.\end{aligned} \tag{A3}$$

Now consider any B with $r(B) \le \hat{r}$,

$$\sum_{i=1}^{h}w_i d_i(XB)-\sum_{i=1}^{h}w_i d_i(X\hat{C}_{\mathrm{S}})=w_{\hat{r}}\sum_{i=1}^{\hat{r}}d_i(XB)-w_{\hat{r}}\sum_{i=1}^{\hat{r}}d_i(X\hat{C}_{\mathrm{S}})+\sum_{i=1}^{\hat{r}}(w_{\hat{r}}-w_i)d_i(X\hat{C}_{\mathrm{S}})-\sum_{i=1}^{\hat{r}}(w_{\hat{r}}-w_i)d_i(XB).$$

Recall that $w_i=d_i^{-\gamma}(PY)$, so $w_r = w_1 \ge \cdots \ge w_r = w_{r-1} \ge 0$. Therefore, both $p_1(\cdot)=\sum_{i=1}^{\hat{r}}d_i(\cdot)$ and $p_2(\cdot)=\sum_{i=1}^{\hat{r}}(w_{\hat{r}}-w_i)d_i(\cdot)$ satisfy the triangle inequality; see the proof of Theorem 1. Moreover, Weyl's inequalities (Franklin, 2000) and the equality $PY = XC_0 + PE$ imply that $d_r(PY) \ge d_r(XC_0) - d_1(PE)$ and $d_1(PY) \le d_1(XC_0) + d_1(PE)$. Hence,

$$\begin{aligned}\sum_{i=1}^{h}w_i d_i(XB)-\sum_{i=1}^{h}w_i d_i(X\hat{C}_{\mathrm{S}}) &\le w_{\hat{r}}\sum_{i=1}^{\hat{r}}d_i(X\hat{C}_{\mathrm{S}}-XB)+\sum_{i=1}^{\hat{r}}(w_{\hat{r}}-w_i)d_i(X\hat{C}_{\mathrm{S}}-XB)\\ &\le \{2d_{\hat{r}}^{-\gamma}(PY)-d_1^{-\gamma}(PY)\}\sum_{i=1}^{\hat{r}}d_i(X\hat{C}_{\mathrm{S}}-XB)\\ &\le [2\{d_{\hat{r}}(XC_0)-d_1(PE)\}^{-\gamma}-\{d_1(XC_0)+d_1(PE)\}^{-\gamma}]\sum_{i=1}^{\hat{r}}d_i(X\hat{C}_{\mathrm{S}}-XB)\\ &\le [2\{d_{\hat{r}}(XC_0)-d_1(PE)\}^{-\gamma}-\{d_1(XC_0)+d_1(PE)\}^{-\gamma}]\,\|X\hat{C}_{\mathrm{S}}-XB\|_F\sqrt{\hat{r}}.\end{aligned}$$

The last inequality is due to the Cauchy–Schwarz inequality. Using (A3), $r(X\hat{C}_S - XB) \le r(\hat{C}_S - B) \le 2r$ ând the inequality $2xy \le x^2/a + ay^2$, we have

$$\begin{aligned}\|X\hat{C}_{\mathrm{S}}-XC_0\|_F^2 &\le \|XB-XC_0\|_F^2+a\|X\hat{C}_{\mathrm{S}}-XB\|_F^2\\ &+\tfrac{1}{a}[d_1(PE)\sqrt{(2\hat{r})}+2\lambda\{d_{\hat{r}}(XC_0)-d_1(PE)\}^{-\gamma}\sqrt{\hat{r}}-\lambda\{d_1(XC_0)+d_1(PE)\}^{-\gamma}\sqrt{\hat{r}}]^2.\end{aligned}$$

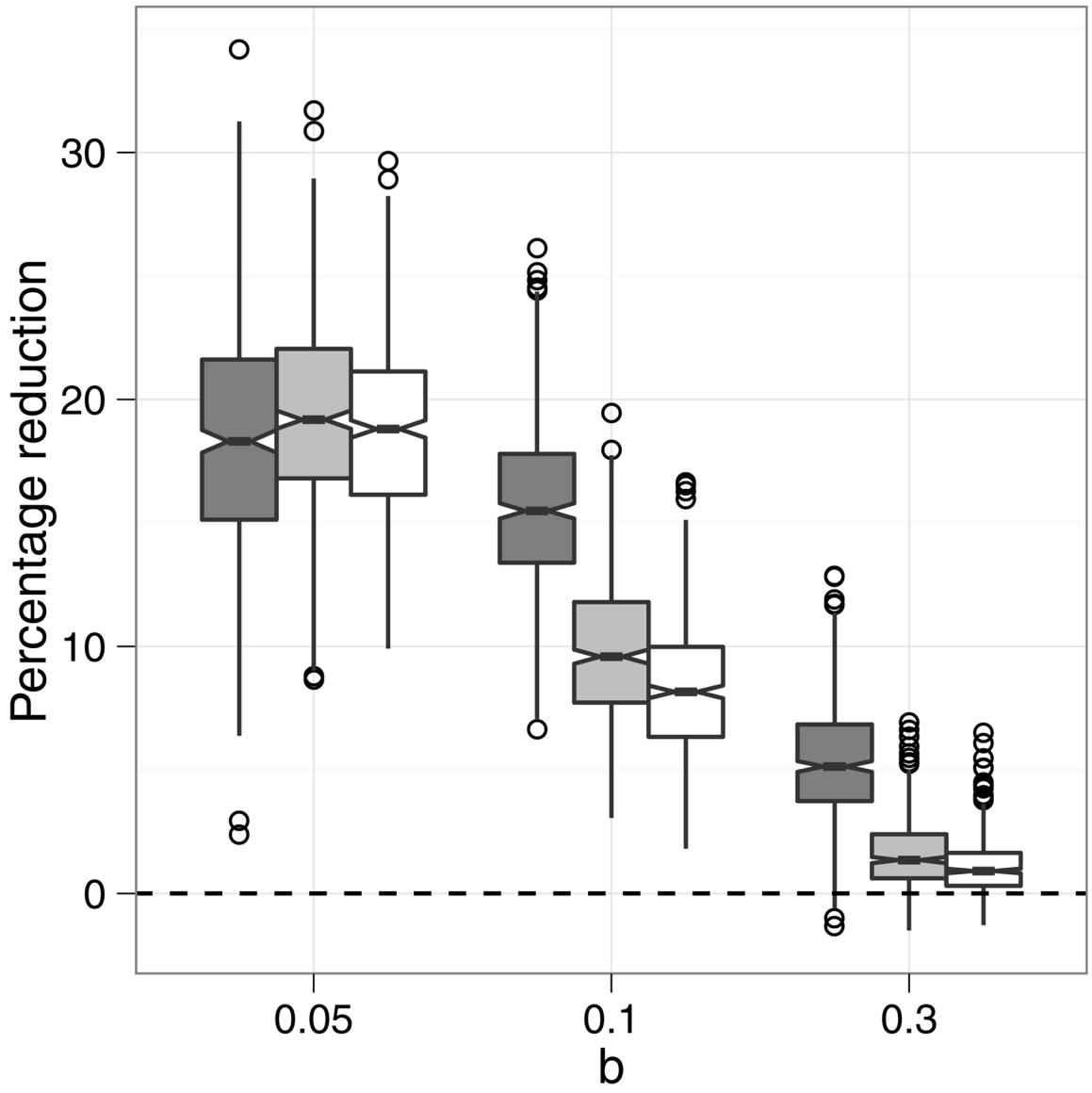
Since $\|X\hat{C}_{\mathrm{S}}-XB\|_F^2 \le \|X\hat{C}_{\mathrm{S}}-XC_0\|_F^2+\|XB-XC_0\|_F^2$, consequently, for any $0 < a < 1$,

$$\begin{aligned}\|X\hat{C}_{\mathrm{S}}-XC_0\|_F^2 &\le \tfrac{1+a}{1-a}\|XB-XC_0\|_F^2\\ &+\tfrac{1}{a(1-a)}[d_1(PE)\sqrt{2}+2\lambda\{d_{\hat{r}}(XC_0)-d_1(PE)\}^{-\gamma}-\lambda\{d_1(XC_0)+d_1(PE)\}^{-\gamma}]^2\hat{r}.\end{aligned}$$

As shown in Theorem 3, on the event $\{d_1(PE) < \delta\lambda^{1/(\gamma+1)}\}$, the estimated rank r êquals the true rank r^*, i.e., $r \hat{=} r^*$, and $\mathrm{pr}\{d_1(PE) \geq \delta\lambda^{1/(\gamma+1)}\} \leq \exp\{-\theta^2(r_x + q)/2\}$. Also, $d_{r^*}(XC_0) > 2\lambda^{1/(\gamma+1)}$ and $c = d_1(XC_0)/d_{r^*}(XC_0) \geq 1$. Therefore, with probability at least $1 - \exp\{-\theta^2(r_x + q)/2\}$,

$$\begin{aligned}\|X\hat{C}_S - XC_0\|_F^2 &\leq \tfrac{1+a}{1-a}\|XB - XC_0\|_F^2 + \tfrac{1}{a(1-a)}\left\{\delta\lambda^{1/(\gamma+1)}\sqrt{2} + 2\lambda(2-\delta)^{-\gamma}\lambda^{-\gamma/(\gamma+1)} - \lambda(2c+\delta)^{-\gamma}\lambda^{-\gamma/(\gamma+1)}\right\}^2 r^* \\ &\leq \tfrac{1+a}{1-a}\|XB - XC_0\|_F^2 + \tfrac{1}{a(1-a)}\{\delta\sqrt{2} + 2(2-\delta)^{-\gamma} - (2c+\delta)^{-\gamma}\}^2\lambda^{2/(\gamma+1)}r^*.\end{aligned}$$

Since B is an arbitrary matrix with $r(B) \leq r^*$, the second part of the theorem is obtained by taking $B = C_0$ and $a = 1/2$. This completes the proof.



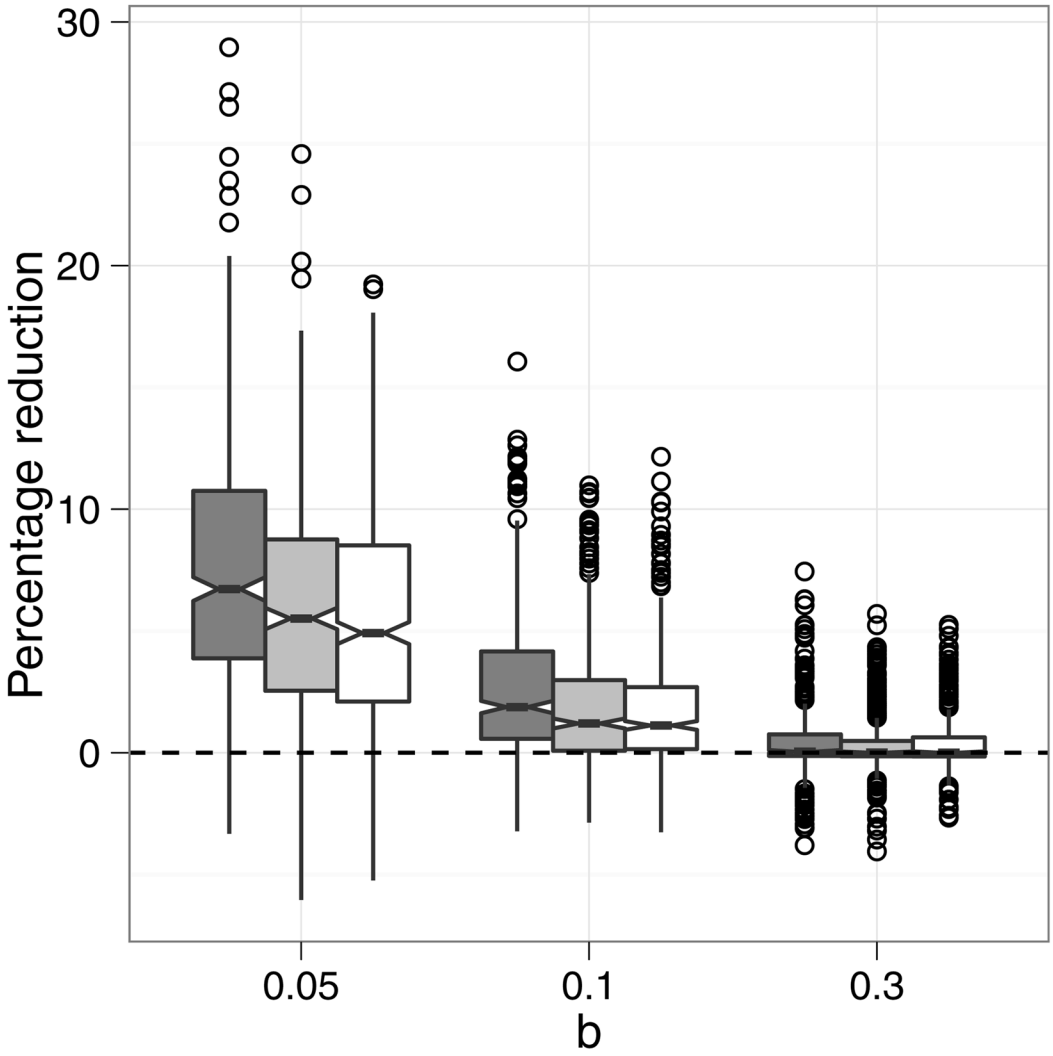
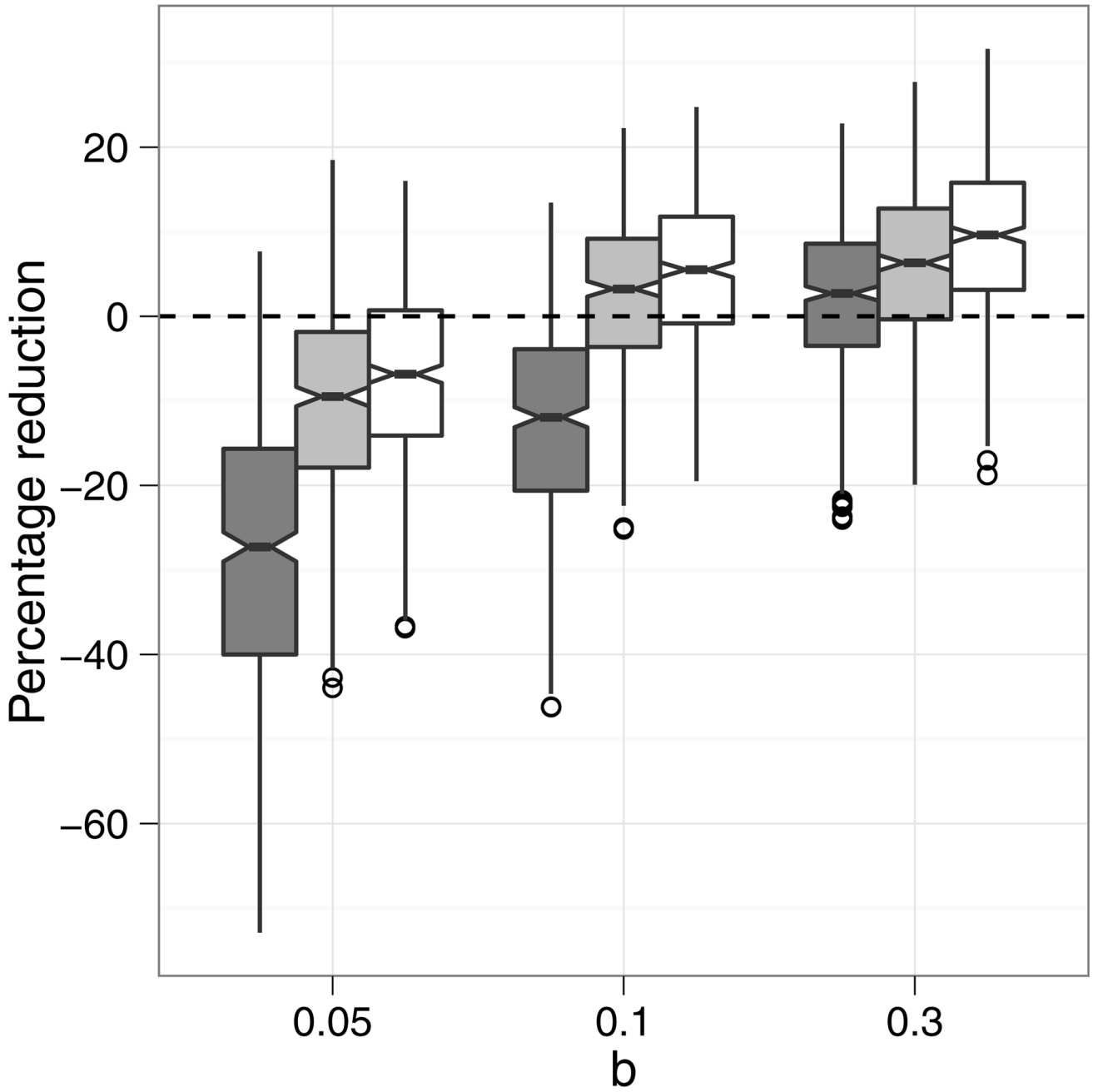


Fig. 1.

Notched boxplots of the percentage reduction in the mean squared prediction error of $\hat{C}_S$ relative to $\hat{C}_H$. The boxplots are shaded in dark grey for $\rho = 0.9$, in light grey for $\rho = 0.5$, and in white for $\rho = 0.1$.



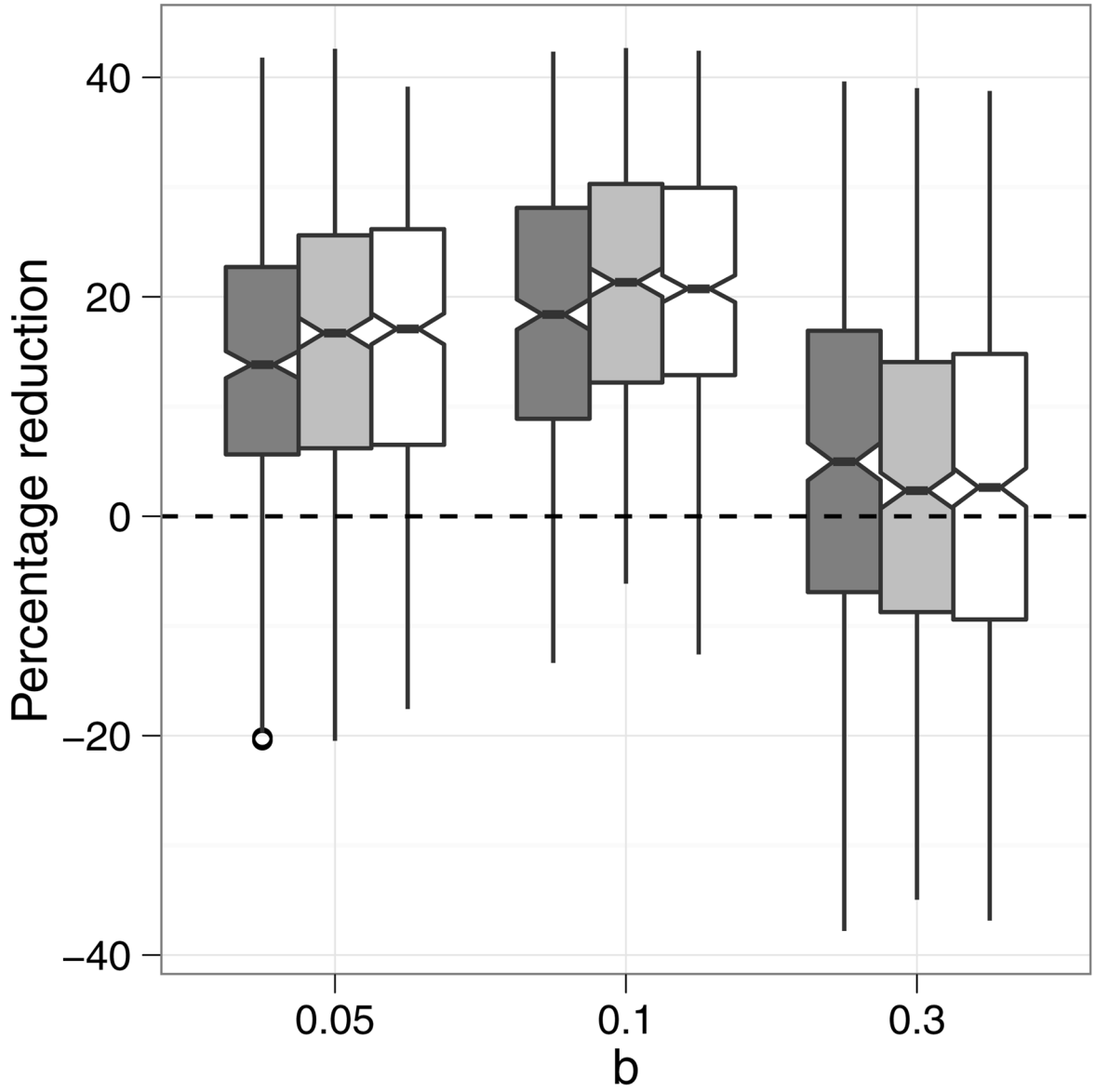


Fig. 2.

Notched boxplots of the percentage reduction in the mean squared prediction error of $\hat{C}_S$ relative to $\hat{C}_N$. All other settings are same as in Figure 1.

Table 1

Comparison of optimally tuned reduced rank estimators using Model I. The estimation and prediction errors are reported, with their standard errors in parentheses.

b		$\hat{C}_N$	$\hat{C}_H$	$\hat{C}_S(2)$	$\hat{C}_S(0)$	$\hat{C}_{HR}$	$\hat{C}_{SR}(2)$
				$\rho = 0{\cdot}9$			
0·05	Est	1·6 (0·2)	3·2 (0·6)	2·6 (0·4)	2·5 (0·3)	1·6 (0·2)	2·5 (0·4)
	Pred	7·8 (0·8)	12·2 (1·3)	9·9 (1·0)	10·5 (1·0)	8·1 (0·8)	9·8 (1·0)
	Rank	7·7, 3%	3·3, 0%	5·5, 0%	10·9, 26%	7·6, 4%	6·0, 0%
0·1	Est	3·5 (0·4)	5·6 (0·7)	4·5 (0·5)	4·3 (0·5)	3·6 (0·4)	4·4 (0·5)
	Pred	12·1 (1·0)	16·1 (1·4)	13·5 (1·2)	14·4 (1·1)	12·5 (1·0)	13·4 (1·1)
	Rank	11·2, 19%	6·2, 0%	8·1, 5%	14·1, 0%	7·9, 5%	8·4, 9%
0·3	Est	6·4 (0·7)	6·8 (0·8)	6·2 (0·7)	6·8 (0·7)	6·1 (0·7)	6·2 (0·7)
	Pred	16·4 (1·2)	16·9 (1·2)	16·0 (1·2)	18·5 (1·2)	15·9 (1·2)	15·9 (1·2)
	Rank	10·7, 30%	9·8, 81%	10·3, 64%	17·0, 0%	9·9, 86%	10·4, 59%
				$\rho = 0{\cdot}5$			
0·05	Est	0·8 (0·1)	1·2 (0·1)	0·9 (0·1)	0·9 (0.1)	0·8 (0·1)	0·9 (0·1)
	Pred	12·4 (1·0)	16·8 (1·5)	13·6 (1·2)	13·3 (1·1)	13·0 (1·1)	13·0 (1·1)
	Rank	12·8, 1%	6·0, 0%	8·0, 5%	13·3, 0%	7·9, 6%	9·1, 27%
0·1	Est	1·2 (0·1)	1·4 (0·1)	1·2 (0·1)	1·2 (0·1)	1·2 (0·1)	1·1 (0·1)
	Pred	16·2 (1·2)	17·4 (1·2)	15·7 (1·1)	16·8 (1·1)	15·7 (1·1)	15·4 (1·1)
	Rank	15·2, 0%	9·2, 32%	10·0, 58%	15·8, 0%	9·4, 47%	10·4, 52%
0·3	Est	1·3 (0·1)	1·3 (0·1)	1·2 (0·1)	1·5 (0·1)	1·2 (0·1)	1·2 (0·1)
	Pred	17·1 (1·3)	16·2 (1·2)	16·0 (1·2)	19·5 (1·3)	16·0 (1·2)	15·9 (1·2)
	Rank	10·9, 14%	10·0, 100%	10·2, 80%	17·8, 0%	10·0, 100%	10·3, 74%
				$\rho = 0{\cdot}1$			
0·05	Est	0·6 (0·1)	0·9 (0·1)	0·7 (0·1)	0·7 (0·1)	0·7 (0·1)	0·7 (0·1)
	Pred	13·3 (1·0)	17·4 (1·4)	14·1 (1·2)	13·7 (1·1)	14·0 (1·2)	13·5 (1·1)
	Rank	14·3, 0%	6·6, 1%	8·5, 12%	13·6, 0%	8·1, 10%	9·6, 35%
0·1	Est	0·9 (0·1)	0·9 (0·1)	0·8 (0·1)	0·9 (0·1)	0·8 (0·1)	0·8 (0·1)

b		$\hat{C}_N$	$\hat{C}_H$	$\hat{C}_S(2)$	$\hat{C}_S(0)$	$\hat{C}_{HR}$	$\hat{C}_{SR}(2)$
	Pred	16·8 (1·2)	17·3 (1·3)	15·9 (1·2)	17·2 (1·2)	16·1 (1·2)	15·5 (1·2)
	Rank	16·5, 0%	9·5, 52%	10·2, 64%	16·1, 0%	9·6, 63%	10·6, 44%
0·3	Est	0·9 (0·1)	0·8 (0·1)	0·8 (0·1)	1·0 (0·1)	0·8 (0·1)	0·8 (0·1)
	Pred	17·7 (1·5)	16·2 (1·2)	16·0 (1·2)	19·6 (1·3)	16·0 (1·2)	16·0 (1·2)
	Rank	11·3, 3%	10·0, 100%	10·2, 83%	17·9, 0%	10·0, 100%	10·3, 76%
	Time	17·5	0·0	0·2	0·2	3·9	2·1

$\hat{C}_N$, nuclear norm penalized estimator (Yuan et al., 2007); $\hat{C}_H$, rank penalized estimator (Bunea et al., 2011); $\hat{C}_S(\gamma)$: adaptive nuclear norm penalized estimator with weight parameter γ; $\hat{C}_{HR}$, robustified rank penalized estimator; $\hat{C}_{SR}(\gamma)$, robustified adaptive nuclear norm penalized estimator with weight parameter γ; Est, estimation error; Pred, prediction error; Rank, average of estimated rank and percentage of correct rank identification; Time, average computation time in seconds per simulation run.

Table 2

Comparison of optimally tuned reduced rank estimators using Model II. The layout of the table is the same as in Table 1.

b		$\hat{C}_N$	$\hat{C}_H$	$\hat{C}_S(2)$	$\hat{C}_S(0)$	$\hat{C}_{HR}$	$\hat{C}_{SR}(2)$
				$\rho = 0{\cdot}9$			
0·05	Est	1·1 (0·2)	1·2 (0·2)	1·2 (0·2)	1·2 (0·2)	1·2 (0·2)	1·2 (0·2)
	Pred	34·1 (4·1)	31·5 (4)	29·1 (3·5)	35·3 (4·2)	29·8 (3·5)	28·9 (3·5)
	Rank	7·4, 2%	4·7, 73%	5·2, 71%	8·3, 0%	4·8, 83%	5·3, 66%
0·1	Est	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)
	Pred	37·0 (4·4)	30·8 (3·6)	30·0 (3·6)	38·9 (4·8)	30·3 (3·6)	29·9 (3·6)
	Rank	7·6, 3%	5·0, 99%	5·2, 79%	8·8, 0%	5·0, 99%	5·3, 75%
0·3	Est	40·4 (5·6)	40·6 (5·9)	40·6 (5·9)	40·6 (5·9)	40·6 (5·9)	40·6 (5·9)
	Pred	32·3 (5·5)	30·2 (3·5)	30·1 (3·5)	41·3 (5·0)	30·0 (3·5)	30·0 (3·5)
	Rank	5·5, 75%	5·0, 100%	5·2, 82%	9·1, 0%	5·0, 100%	5·3, 75%
				$\rho = 0{\cdot}5$			
0·05	Est	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)
	Pred	35·0 (4·2)	31·0 (3·9)	29·1 (3·6)	35·9 (4·7)	29·7 (3·6)	29·0 (3·6)
	Rank	8·0, 0%	4·9, 90%	5·2, 77%	8·4, 0%	4·9, 93%	5·3, 68%
0·1	Est	4·5 (0·6)	4·5 (0·6)	4·5 (0·6)	4·5 (0·6)	4·5 (0·6)	4·5 (0·6)
	Pred	37·8 (4·6)	30·2 (3·5)	29·6 (3·3)	39·0 (4·2)	29·8 (3·4)	29·5 (3·4)
	Rank	7·8, 7%	5·0, 99%	5·2, 81%	8·9, 0%	5·0, 99%	5·3, 74%
0·3	Est	40·4 (6·1)	40·1 (6·0)	40·1 (6·0)	40·1 (6·0)	40·1 (6·0)	40·1 (6·0)
	Pred	31·6 (4·9)	30·3 (3·1)	30·2 (3·1)	41·3 (4·0)	30·2 (3·1)	30·2 (3·1)
	Rank	5·3, 86%	5·0, 100%	5·2, 78%	9·2, 0%	5·0, 100%	5·3, 70%
				$\rho = 0{\cdot}1$			
0·05	Est	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)	1·1 (0·2)
	Pred	35·2 (4·2)	31·0 (3·8)	29·3 (3·5)	36·0 (4·3)	29·9 (3·5)	29·1 (3·6)
	Rank	8·1, 0%	4·9, 88%	5·2, 76%	8·4, 0%	4·9, 91%	5·3, 65%
0·1	Est	4·6 (0·7)	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)	4·5 (0·7)
	Pred	37·7 (4·6)	30·2 (3·5)	29·7 (3·5)	38·9 (4·4)	29·8 (3·5)	29·6 (3·4)

b		$\hat{C}_N$	$\hat{C}_H$	$\hat{C}_S(2)$	$\hat{C}_S(0)$	$\hat{C}_{HR}$	$\hat{C}_{SR}(2)$
0·3	Rank	7·9, 5%	5·0, 100%	5·2, 79%	8·8, 0%	5·0, 100%	5·3, 73%
	Est	40·5 (6·2)	40·5 (5·6)	40·5 (5·6)	40·5 (5·6)	40·5 (5·6)	40·5 (5·6)
	Pred	31·3 (5·1)	30·1 (3·4)	30·0 (3·4)	41·2 (4·6)	29·9 (3·4)	29·9 (3·4)
	Rank	5·3, 86%	5·0, 100%	5·2, 76%	9·1, 0%	5·0, 100%	5·3, 68%
Time		20·2	0·0	0·2	0·2	4·0	2·5

Table 3

Comparison of the model fits to the chromosome 21 data for various reduced rank methods. The mean squared prediction errors and the estimated ranks are reported, with their standard errors in parentheses.

	$\hat{C}_N$	$\hat{C}_H$	$\hat{C}_{HR}$	$\hat{C}_S(2)$	$\hat{C}_{SR}(2)$
			Full data		
MSPE	0·71 (0·2)	0·.69 (0·1)	0·68 (0·1)	0·68 (0·1)	0·68 (0.·1)
Rank	6·2 (0·8)	1·0 (0·0)	1·0 (0·0)	1·0 (0·2)	1·8 (0·4)
			Selected predictors		
MSPE	0·85 (0·2)	0·90 (0·2)	0·82 (0·1)	0·85 (0·1)	0·84 (0·1)
Rank	4·6 (0·7)	1·3 (0·5)	1·9 (0·3)	2·1 (0·2)	2·4 (0·5)

MSPE, average of mean squared prediction errors based on 100 replications of 79/10 splits of the data; Rank: corresponding average of the rank estimates. The notations of the estimators are the same as in Table 1.