Biometrika (2015), **102**, 2, *pp.* 457–477
Printed in Great Britain

doi: 10.1093/biomet/asu067
Advance Access publication 9 February 2015

On the degrees of freedom of reduced-rank estimators in multivariate regression

By A. MUKHERJEE

Smart Forecasting Team, @WalmartLabs, 850 Cherry Avenue, San Bruno, California 94066, U.S.A.

ashinm@umich.edu

K. CHEN

Department of Statistics, University of Connecticut, 215 Glenbrook Road U-4120, Storrs, Connecticut 06269, U.S.A.

kun.chen@uconn.edu

N. WANG and J. ZHU

Department of Statistics, University of Michigan, 1085 S. University Avenue, Ann Arbor, Michigan 48109, U.S.A.

nwangaa@umich.edu jizhu@umich.edu

Summary

We study the effective degrees of freedom of a general class of reduced-rank estimators for multivariate regression in the framework of Stein's unbiased risk estimation. A finite-sample exact unbiased estimator is derived that admits a closed-form expression in terms of the thresholded singular values of the least-squares solution and hence is readily computable. The results continue to hold in the high-dimensional setting where both the predictor and the response dimensions may be larger than the sample size. The derived analytical form facilitates the investigation of theoretical properties and provides new insights into the empirical behaviour of the degrees of freedom. In particular, we examine the differences and connections between the proposed estimator and a commonly-used naive estimator. The use of the proposed estimator leads to efficient and accurate prediction risk estimation and model selection, as demonstrated by simulation studies and a data example.

Some key words: Adaptive nuclear norm; Degrees of freedom; Model selection; Multivariate regression; Reduced-rank regression; Singular value decomposition.

1. Introduction

Multivariate linear regression extends the classical univariate regression model to $q > 1$ responses and p predictors. It is commonly used in bioinformatics, chemometrics, econometrics, and other fields where one is interested in predicting several responses simultaneously. Let X denote the $n \times p$ predictor or design matrix and Y the $n \times q$ response matrix. The regression parameters are given by the $p \times q$ coefficient matrix B. The kth column of B is the regression coefficient vector for regressing the kth response on the predictors. Let ϵ denote the $n \times q$ random error matrix with independent entries having mean zero and variance σ^2. Then the multivariate linear regression model is

$$Y = XB + \epsilon. \tag{1}$$

This reduces to the classical univariate regression model when $q = 1$. For notational simplicity we assume that the responses and predictors are centred, and hence the intercept term can be omitted without loss of generality. We assume p and q to be fixed but unrestricted by the sample size n. When the Gram matrix $X^{\mathrm{T}}X$ is invertible, the ordinary least-squares approach to estimating B leads to $\hat{B}_{\text{ols}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. It amounts to performing q separate univariate regressions and completely ignores the multivariate aspect of the problem, where the responses may be highly correlated and hence the effective dimensionality can be much smaller than q. Also, least squares is unsuitable for the high-dimensional case where both p and q are greater than n. Many methods have been proposed to overcome these drawbacks, under the general class of linear factor regression, where the responses are regressed against a small number of linear combinations of predictors, commonly known as factors. Examples include principal components regression (Massy, 1965), partial least squares (Wold, 1975) and canonical correlation analysis (Hotelling, 1935). These methods differ in the way they choose the factors. Recently, Witten et al. (2009) introduced a penalized canonical correlation analysis using sparse matrix factorization that leads to more interpretable factors and is more suitable for high-dimensional problems. Breiman & Friedman (1997) proposed a two-step approach which borrows strength by performing a second round of regression of the responses on the ordinary least-squares estimators, and they showed connections of this approach with canonical correlation analysis.

Yet another line of research focuses on the rank of the regression coefficient matrix. Anderson (1951) proposed a class of regression models that restrict the rank of the coefficient matrix to be much smaller than the dimensionality of B; that is, $\text{rank}(B) \leqslant r \leqslant \min(p, q)$. This is reasonable in many multivariate regression problems, and can be interpreted as follows: the q responses are related to the p predictors only through r effective linear factors, leading to the optimization problem

$$\hat{B}(r) = \underset{\{B:\, \text{rank}(B) \leqslant r\}}{\arg\min} \|Y - XB\|_{\mathrm{F}}^2, \tag{2}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. Even though the rank constraint makes (2) a nonconvex optimization problem, it admits a closed-form solution, as we shall see later. Izenman (1975) introduced the term reduced-rank regression for this class of models and derived the asymptotic distributions and confidence intervals for such estimators. A non-exhaustive list of notable work includes Rao (1978), Davies & Tso (1982) and Anderson (1999, 2002b); see Reinsel & Velu (1998) and Izenman (2008) for more comprehensive accounts. Recently, there has been a revival of interest in reduced-rank methods. Instead of restricting the rank, Yuan et al. (2007) proposed putting an ℓ_1 penalty on the singular values of B, also known as the nuclear norm. The nuclear-norm-penalized least-squares criterion encourages sparsity among the singular values to achieve simultaneous rank reduction and shrinkage coefficient estimation (Neghaban & Wainwright, 2011; Lu et al., 2012), but it is computationally intensive and tends to overestimate the rank. Bunea et al. (2011) proposed a rank selection criterion that extends reduced-rank regression to high-dimensional settings, in which rank-constrained estimation was cast as a penalized least-squares method with the penalty proportional to the rank of the coefficient matrix or, equivalently, the ℓ_0-norm of its singular values. Chen et al. (2012) adopted sparsity penalties on singular vectors for reduced-rank regression problems, leading to more interpretable latent model structures. Bunea et al. (2012) and Chen & Huang (2012) suggested imposing row sparsity on the reduced-rank coefficient matrix for conducting predictor selection. Chen et al. (2013) proposed an adaptive nuclear-norm penalty on the signal matrix XB to close the gap between ℓ_0 and ℓ_1 penalties on singular values.

In this paper, we study the degrees of freedom of reduced-rank estimators in multivariate linear regression models. Degrees of freedom is one of the most widely used terms in statistics, but it has largely been overlooked in reduced-rank regression except for some heuristic suggestions (Davies & Tso, 1982; Reinsel & Velu, 1998). For example, the number of free parameters in a $p \times q$ matrix of rank r, $(p+q-r)r$, has been suggested as a naive estimate of the degrees of freedom of the reduced-rank regression estimator when restricted to rank $r \leqslant \min(p, q)$. More precisely, for an arbitrary design matrix, the number of free parameters should be $(r_x + q - r)r$, where $r_x = \text{rank}(X)$ is the rank of the design matrix (Bunea et al., 2011). Henceforth, we refer to this as the naive estimator of the degrees of freedom of a rank-r model. In this paper, we develop a finite-sample unbiased estimator of the degrees of freedom for a general class of reduced-rank estimators for the multivariate regression model and investigate its properties. Our results are nonasymptotic, so the estimator is valid for any given model dimensions and sample size.

In a nutshell, the degrees of freedom quantifies the complexity of a statistical modelling procedure (Hastie & Tibshirani, 1990). In the case of the univariate linear regression model, the degrees of freedom is the number of estimated parameters, p. However, in general there is no exact correspondence between the degrees of freedom and the number of free parameters in the model (Ye, 1998). For example, in best-subset selection for univariate regression (Hocking & Leslie, 1967), we search for the best model of size $p_0 \in \{1, 2, \ldots, p\}$ that minimizes the residual sum of squares. The resulting model has p_0 parameters, but intuitively the degrees of freedom would be higher than p_0, since the search for the optimal subset of size p_0 increases model complexity (Hastie et al., 2009). In other words, for best-subset selection the optimal p_0-dimensional subspace that minimizes the residual sum of squares clearly depends on Y. Thus, the final estimator is highly nonlinear in Y, which results in the loss of correspondence between the degrees of freedom and the number of parameters in the model.

Similar arguments apply to reduced-rank regression. Instead of searching for the best p_0 variables, as in best-subset selection, here we are searching for the best r linear combinations of the predictors that minimize the squared loss, which should intuitively suggest increased model complexity. Since the optimal rank-r subspace depends on the response matrix Y, the correspondence between the number of free parameters and the degrees of freedom need not hold. Thus reduced-rank regression is different from other linear factor regression methods such as principal components regression (Massy, 1965). In principal components regression, the factors are principal components of the design matrix X, which do not depend on the response Y, and so the final estimator is still linear in Y.

2. Degrees of freedom

Stein (1981), in his theory of unbiased risk estimation, first introduced a rigorous definition of the degrees of freedom of a statistical estimation procedure. Later, Efron et al. (2004) showed that Stein's treatment can be considered a special case of a more general notion under the assumption of Gaussianity. Assume that we have data of the form $(y_{n\times 1}, X_{n\times p})$. Given X, the response originates from the model $y \sim (\mu, \sigma^2 I)$ where μ is the true mean, which can be a function of X, and σ^2 is the common variance. Then, for any estimation procedure $m(\cdot)$ with fitted values $\hat{\mu} = m(X, y)$, the degrees of freedom of $m(\cdot)$ is defined as

$$\text{df}(m) = \sum_{i=1}^{n} \text{cov}(\hat{\mu}_i, y_i)/\sigma^2. \tag{3}$$

The rationale is that more complex models would try to fit the data better, and hence the covariance between observed and fitted pairs would be higher. This expression is not directly observable except in certain simple cases, such as when $m(y) = Sy$, a linear smoother; in that case, it is not difficult to see that $\mathrm{df}(m) = \mathrm{tr}(S)$, which agrees with the usual definition of degrees of freedom (Hastie & Tibshirani, 1990). Stein was able to overcome this hurdle for a special case where $y \sim N(\mu, \sigma^2 I)$. Using a simple equality for the Gaussian distribution, he proved that as long as the partial derivative $\partial \hat{\mu}_i / \partial y_i$ exists almost everywhere for all $i \in \{1, \ldots, n\}$,

$$\mathrm{cov}(\hat{\mu}_i, y_i) = \sigma^2 E\left(\frac{\partial \hat{\mu}_i}{\partial y_i}\right),$$

giving the following unbiased estimator of the degrees of freedom for the fitting procedure $m(\cdot)$:

$$\hat{\mathrm{df}}(m) = \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}. \tag{4}$$

Using the definition of degrees of freedom in (3), Efron et al. (2004) employed a covariance penalty approach to prove that the C_P-type statistics of Mallows (1973) provide an unbiased estimator of the true prediction error, namely

$$C_P(\hat{\mu}) = \frac{1}{n}\|y - \hat{\mu}\|^2 + \frac{2\,\mathrm{df}(\hat{\mu})}{n}\sigma^2.$$

This reveals the important role played by the degrees of freedom in model assessment, provides a principled way to select the optimal model without using computationally expensive methods such as crossvalidation, and can in certain settings offer significantly better prediction accuracy than those computationally expensive methods (Efron et al., 2004). Indeed, the degrees of freedom is an integral part of almost every popular model selection criterion, including the Bayesian information criterion (Schwarz, 1978) and generalized crossvalidation (Golub et al., 1979). Many important works followed Stein (1981) and Efron et al. (2004). Donoho & Johnstone (1995) used the unbiased risk estimation framework to derive the degrees of freedom for the soft-thresholding operator in wavelet shrinkage; Meyer & Woodroofe (2000) employed this framework to derive the same for shape-restricted regression; and Li & Zhu (2008) used this approach to compute an unbiased estimator of the degrees of freedom for penalized quantile regression. Zou et al. (2007) applied Stein's theory of unbiased risk estimation to the lasso (Tibshirani, 1996); this is challenging due to the nonlinear nature of the lasso solution, which does not admit an analytical solution except in some special cases. Using sophisticated mathematical analysis, Zou et al. (2007) were able to prove that the number of nonzero coefficients provides an unbiased estimator of the degrees of freedom for the lasso. This result is of great practical importance, as it allows one to use model selection criteria such as C_P or BIC for the lasso without extra computational cost.

The degrees of freedom for reduced-rank estimators also presents challenges because of the nonlinearity of the estimator. Even though it admits a closed-form solution, the solution is highly nonlinear, depending on the singular value decomposition of the least-squares solution, to be described in (5). Below we study the degrees of freedom of a general class of reduced-rank estimators in the framework of unbiased risk estimation and propose a finite-sample exact unbiased estimator. The importance of such an estimator has been emphasized by Shen & Ye (2002), Efron et al. (2004), Zou et al. (2007) and others.

To overcome the analytical difficulties in computing the degrees of freedom, Ye (1998) and Shen & Ye (2002) proposed the generalized degrees-of-freedom approach, where they evaluate the partial derivatives in (4) numerically, using data-perturbation techniques to compute an approximately unbiased estimator. Efron et al. (2004) used a parametric bootstrap to arrive at an approximately unbiased estimator of (3). In their method, the objective is to directly estimate $\mathrm{cov}(\hat{y}_i, y_i)$ by drawing repeated samples from the underlying distribution and fitting the model. In the absence of such samples, this can be achieved by using a parametric bootstrap to simulate data from a larger unbiased model and computing the covariance between the fitted and observed values. Although these simulation approaches allow one to extend the degrees-of-freedom approach to many highly nonlinear modelling frameworks, they are computationally expensive, and the lack of a closed-form expression makes investigation of the theoretical properties difficult.

3. A CLASS OF REDUCED-RANK ESTIMATORS

Recall the multivariate linear regression model in (1). Let $\hat{Y}$ be the least-squares estimate, which admits a singular value decomposition of the form

$$\hat{Y} = X(X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y = \underset{n\times\bar{r}}{W}\ \underset{\bar{r}\times\bar{r}}{D}\ \underset{\bar{r}\times q}{V^{\mathrm{T}}}, \tag{5}$$

where A^{-} denotes the Moore–Penrose pseudo-inverse of a generic matrix A (Moore, 1920; Penrose, 1955). The dimensions p and q are assumed to be fixed but are not restricted by the sample size n. The Moore–Penrose pseudo-inverse is well-defined for an arbitrary choice of (p, q, n) as well as for a rank-deficient design matrix X. In (5), W and V are orthogonal matrices that represent the left and right singular vectors, and $D = \mathrm{diag}\{d_i : i = 1, \ldots, \bar{r}\}$ where $d_1 \geqslant \cdots \geqslant d_{\bar{r}} > 0$ are the nonzero singular values of $\hat{Y}$. Without loss of generality we assume that $\mathrm{rank}(\hat{Y}) = \bar{r} = \min(r_x, q)$, where r_x denotes the rank of the design matrix. We will denote the kth columns of W and V by w_k and v_k, respectively. Using the Eckart–Young theorem (Eckart & Young, 1936), it is not difficult to show that the reduced-rank regression estimator for (2) can be expressed as

$$\hat{Y}(r) = \hat{Y}\sum_{k=1}^{r} v_k v_k^{\mathrm{T}} = W^{(r)}D^{(r)}V^{(r)\mathrm{T}} \quad (r = 1, \ldots, \bar{r}), \tag{6}$$

where $A^{(r)}$ denotes the first r columns of a generic matrix A. This rank-constrained estimation procedure can also be viewed under a more general penalized least-squares framework,

$$\min_{B}\left\{\frac{1}{2}\|Y - XB\|_{\mathrm{F}}^{2} + \lambda\mathcal{P}(B)\right\}, \tag{7}$$

in which the penalty is proportional to the rank of the coefficient matrix B, i.e., $\mathcal{P}(B) = \mathrm{rank}(B)$ (Bunea et al., 2011). It leads to a hard-thresholding of the singular values of $\hat{Y}$. More generally, under the regularized estimation framework (7), a set of reduced-rank estimators can be indexed by the regularization parameter λ, which controls the penalty level and hence the model's complexity. In light of that, we consider a broad class of such reduced-rank estimators, defined by

$$\tilde{Y}(\lambda) = X\tilde{B}(\lambda) = \sum_{k=1}^{\bar{r}} s_k(d_k, \lambda) d_k w_k v_k^{\mathrm{T}} = \hat{Y}\sum_{k=1}^{\bar{r}} s_k(d_k, \lambda) v_k v_k^{\mathrm{T}}, \tag{8}$$

where each $s_k(d_k, \lambda) \in [0, 1]$ is a function of d_k and λ, such that $s_1(d_1, \lambda) \geqslant \cdots \geqslant s_{\bar{r}}(d_{\bar{r}}, \lambda) \geqslant 0$. For simplicity, we write $s_k(d_k, \lambda) = s_k(\lambda) = s_k$. The reduced-rank regression estimator can be viewed as a special case of this general framework, with $s_k(d_k, r) = \mathbb{1}(k \leqslant r) \in \{0, 1\}$ $(r = 1, \ldots, \bar{r})$, where the solutions are indexed by the rank constraint r instead of a continuous penalty parameter λ. This class of estimators has the same set of singular vectors as the reduced-rank regression estimator in (6), but may have different singular value estimates given by shrunk or thresholded versions of the estimated singular values from least squares. Such estimators can be obtained from a nonconvex singular value penalization or from thresholding operations (She, 2009, 2013; Chen et al., 2013). The class of estimators (8) is computationally efficient and possesses many desirable theoretical properties, such as rank selection consistency and attainment of the minimax error bound (Bunea et al., 2011) in both the classical and the high-dimensional regimes. Examples include the rank selection estimator (Bunea et al., 2011), the nuclear-norm-penalized estimator under an orthogonal design (Yuan et al., 2007), and the adaptive nuclear-norm estimator (Chen et al., 2013).

4. DEGREES OF FREEDOM OF REDUCED-RANK ESTIMATORS

In the previous section we discussed a broad class of reduced-rank estimators covering both hard- and soft-thresholding of the singular values of $\hat{Y}$. Next, we apply definition (4) to such multivariate regression estimators to estimate the degrees of freedom. We start by rewriting the multivariate linear regression model (1) as

$$\underset{nq\times 1}{\mathrm{vec}(Y)} = \underset{nq\times pq}{(I_q \otimes X)}\,\underset{pq\times 1}{\mathrm{vec}(B)} + \underset{nq\times 1}{\mathrm{vec}(\epsilon)},$$

where $\otimes$ denotes the usual Kronecker product between matrices and $\mathrm{vec}(\cdot)$ stands for the columnwise vectorization operator on a matrix. We will first derive the results for the special case of the reduced-rank regression estimator (6) and later extend them to the general class of model (8). Applying definition (4), we get

$$\hat{\mathrm{df}}(r) = \mathrm{tr}\left[\frac{\partial\,\mathrm{vec}\{\hat{Y}(r)\}}{\partial\,\mathrm{vec}(Y)}\right] \quad (r = 1, \ldots, \bar{r}), \tag{9}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator for a real square matrix.

Despite its simplicity, direct computation of $\hat{\mathrm{df}}(r)$ from (9) remains difficult. We now show that the problem boils down to determining the divergence measure of a low-rank matrix approximation to a full-rank matrix, regardless of the model dimensionality. Recall that we have assumed $\mathrm{rank}(\hat{Y}) = \bar{r} = \min(r_x, q)$, which is not restrictive in general and does not depend on the dimensions of the problem. Let $X^{\mathrm{T}}X = QS^2Q^{\mathrm{T}}$ be the eigendecomposition of $X^{\mathrm{T}}X$; that is, $Q \in \mathbb{R}^{p\times r_x}$ with $Q^{\mathrm{T}}Q = I$, and $S \in \mathbb{R}^{r_x\times r_x}$ is a diagonal matrix with positive diagonal elements. Then, the Moore–Penrose inverse of $X^{\mathrm{T}}X$ can be written as $(X^{\mathrm{T}}X)^- = QS^{-2}Q^{\mathrm{T}}$. Define

$$H = S^{-1}Q^{\mathrm{T}}X^{\mathrm{T}}Y.$$

It follows that $H \in \mathbb{R}^{r_x\times q}$ admits a singular value decomposition of the form

$$H = UDV^{\mathrm{T}},$$

where $U \in \mathbb{R}^{r_x\times \bar{r}}$ with $U^{\mathrm{T}}U = I$, and V and D are as defined in (5). The matrix H has the same set of singular values and right singular vectors as $\hat{Y}$ in (5), because $H^{\mathrm{T}}H = \hat{Y}^{\mathrm{T}}\hat{Y} = Y^{\mathrm{T}}X(X^{\mathrm{T}}X)^-X^{\mathrm{T}}Y$. Moreover, H is of full rank since $\hat{Y}$ is of rank $\bar{r} = \min(r_x, q)$. The matrix H plays a key role in the derivation of a simple form for the degrees of freedom. In particular, this construction allows us to avoid singularities arising from $r_x < p$ in the high-dimensional scenario.

Upon simplifying (9) using matrix equalities, we obtain

$$\hat{\mathrm{df}}(r) = \mathrm{tr}\left\{ \frac{\partial\, \mathrm{vec}(U^{(r)} D^{(r)} V^{(r)\mathrm{T}})}{\partial\, \mathrm{vec}(H)} \right\} = \mathrm{tr}\left[\frac{\partial\, \mathrm{vec}\{H(r)\}}{\partial\, \mathrm{vec}(H)} \right] = \sum_{i=1}^{r_x} \sum_{j=1}^{q} \frac{\partial h_{ij}(r)}{\partial h_{ij}}, \tag{10}$$

where $H(r) = U^{(r)} D^{(r)} V^{(r)\mathrm{T}} = \{h_{ij}(r)\}_{r_x \times q}$ is the rank-r approximation to H. The details of this derivation are given in the Appendix.

For the general class of reduced-rank estimators in (8), we have

$$\tilde{Y}(\lambda) = XQS^{-1} H \sum_{k=1}^{\bar{r}} s_k(d_k, \lambda) v_k v_k^{\mathrm{T}} = XQS^{-1} U \tilde{D}(\lambda) V^{\mathrm{T}},$$

where $\tilde{D}(\lambda) = \mathrm{diag}\{s_k(d_k, \lambda) d_k : k = 1, \ldots, \bar{r}\}$. Once again, by using familiar matrix algebra, we arrive at a simpler expression for the degrees of freedom for the general class of reduced-rank models:

$$\tilde{\mathrm{df}}(\lambda) = \mathrm{tr}\left[\frac{\partial\, \mathrm{vec}\{U \tilde{D}(\lambda) V^{\mathrm{T}}\}}{\partial\, \mathrm{vec}(H)} \right] = \mathrm{tr}\left[\frac{\partial\, \mathrm{vec}\{\tilde{H}(\lambda)\}}{\partial\, \mathrm{vec}(H)} \right], \tag{11}$$

where $\tilde{H}(\lambda) = U \tilde{D}(\lambda) V^{\mathrm{T}}$.

It is now clear that the problem reduces to computing the divergence of a low-rank approximation of the matrix H with respect to H itself. Such a computation would involve the derivatives of singular values and singular vectors, which are not only highly nonlinear functions of the underlying matrix but also discontinuous on certain subsets of matrices (O'Neil, 2005). This makes calculation of the degrees of freedom for the reduced-rank regression challenging. Stein (1981) used derivatives of the singular values of a positive-semidefinite matrix to estimate the risk improvement for a class of estimators for the mean of a multivariate Gaussian distribution. Tsukuma (2008) used a similar method to prove minimaxity of Bayes estimators for the mean matrix of a Gaussian distribution. Our set-up is very different from those of Stein (1981) and Tsukuma (2008). Specifically, we consider a regression setting where the design matrix makes the derivation more difficult. Also, as we aim to estimate the degrees of freedom of the model, we need the derivatives of both singular values and vectors to compute the right-hand side of (11). A considerable amount of work has been done on the smoothness and differentiability of the singular value decomposition of a real matrix; see Magnus & Neudecker (1998) and O'Neil (2005). In view of this, we proceed in two main steps: (i) derive the partial derivatives in (10) and (11) for the case where H does not have repeated singular values, i.e., $d_1 > d_2 > \cdots > d_{\bar{r}} > 0$, and use them to obtain an explicit exact unbiased estimator of the degrees of freedom; (ii) prove that the set where the partial derivatives do not exist has zero Lebesgue measure.

5. Proposed estimator

We start by examining the derivatives of the singular values and singular vectors of a matrix with respect to an entry of the matrix itself. All the proofs are given in the Appendix.

THEOREM 1. *Suppose that H is a $r_x \times q$ matrix of rank q, with $r_x \geqslant q$. Let its singular value decomposition be $H = UDV^{\mathrm{T}}$, where $U \in \mathbb{R}^{r_x \times q}$ with $U^{\mathrm{T}}U = I$, $V \in \mathbb{R}^{q \times q}$ with $V^{\mathrm{T}}V = I$, and*

$D = \mathrm{diag}\{d_i : i = 1, \ldots, q\}$ with $d_1 > \cdots > d_q > 0$. Then, for each $1 \leqslant i \leqslant r_x$, $1 \leqslant j \leqslant q$ and $1 \leqslant k \leqslant q$,

$$\frac{\partial v_k}{\partial h_{ij}} = -(H^{\mathrm{T}}H - d_k^2 I)^{-}(H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H)v_k,$$

$$\frac{\partial d_k}{\partial h_{ij}} = \frac{1}{2d_k}\, v_k^{\mathrm{T}}(H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H)v_k,$$

where $(H^{\mathrm{T}}H - d_k^2 I)^{-} = V(D^2 - d_k^2 I)^{-}V^{\mathrm{T}}$, with $(\cdot)^{-}$ denoting the Moore–Penrose inverse, and $Z^{(ij)} = \partial H/\partial h_{ij}$ is a $r_x \times q$ matrix of zeros for entry (i, j).

Without loss of generality, we have assumed $r_x \geqslant q$ in the above theorem. When $r_x \leqslant q$, the same results could be stated for H^{T} with r_x and q interchanged. Theorem 1 can be established from the general results in Magnus & Neudecker (1998) about the derivatives of a generalized eigensystem. To ensure that the derivatives are well-defined, we have assumed that the singular values are distinct. This is hardly a restriction in real applications, as the observed singular values rarely coincide. The following theorem states that matrices of full rank and with nonrepeated singular values are dense in the set of all real matrices of dimension $r_x \times q$.

Theorem 2. *Let $\mathbb{R}^{r_x \times q}$ be the space of all real-valued $(r_x \times q)$-dimensional matrices equipped with the Lebesgue measure μ. Also, let $\mathcal{S} \subseteq \mathbb{R}^{r_x \times q}$ denote the subset of matrices that are of full rank and have no repeated singular values. Then $\mu(\mathcal{S}) = 1$.*

It is not immediately clear whether the derived unbiased estimators in (10) and (11) admit an explicit form. Examining the singular value decomposition of H sheds light on this. The pairs of singular vectors (u_k, v_k) are orthogonal to each other, representing distinct directions in $\mathbb{R}^{r_x \times q}$ without any redundancy. Intuitively, these directions are themselves indistinguishable, and their relative importances in constituting the matrix H are entirely revealed by the singular values. This suggests that the complexity of reduced-rank estimation, as reflected by the relative complexity of a low-rank approximation $H(r)$ or $\tilde{H}(\lambda)$ with respect to H, may depend only on the singular values of the matrix H and the mechanism of singular value shrinkage or thresholding. This is the main intuition that motivated our results on explicit forms for (10) and (11), which are summarized in the following two theorems.

Theorem 3. *Let $\hat{Y}$ be the least-squares estimator in (5). Let $r_x = \mathrm{rank}(X)$ and $\bar{r} = \mathrm{rank}(\hat{Y}) = \min(r_x, q)$. Suppose that the singular values of $\hat{Y}$ satisfy $d_1 > \cdots > d_{\bar{r}} > 0$. Consider the reduced-rank estimator $\hat{Y}(r)$ in (6). An unbiased estimator of the effective degrees of freedom is*

$$\hat{\mathrm{df}}(r) = \begin{cases} \max(r_x, q)r + \displaystyle\sum_{k=1}^{r}\sum_{l=r+1}^{\bar{r}} \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2}, & r < \bar{r}, \\ r_x q, & r = \bar{r}. \end{cases}$$

The results are further generalized to the class of reduced-rank estimators in (8). The weights $s_k(d_k, \lambda)$ are treated as random quantities since they are usually functions of the singular values.

Theorem 4. *Let $\hat{Y}$ be the least-squares estimator in (5). Let $\bar{r} = \mathrm{rank}(\hat{Y}) = \min(r_x, q)$, and suppose that the singular values of $\hat{Y}$ satisfy $d_1 > \cdots > d_{\bar{r}} > 0$. Consider the reduced-rank estimator $\tilde{Y}(\lambda)$ in (8), and let $\tilde{r} = \tilde{r}(\lambda) = \max\{k : s_k(d_k, \lambda) > 0.\}$. An unbiased estimator of the* *effective degrees of freedom is*

$$\tilde{\mathrm{df}}(\lambda) = \begin{cases} \max(r_x, q) \sum_{k=1}^{\tilde{r}} s_k + \sum_{k=1}^{\tilde{r}} \sum_{l=\tilde{r}+1}^{\bar{r}} \frac{s_k(d_k^2 + d_l^2)}{d_k^2 - d_l^2} + \sum_{k=1}^{\tilde{r}} \sum_{l \neq k}^{\tilde{r}} \frac{d_k^2(s_k - s_l)}{d_k^2 - d_l^2} + \sum_{k=1}^{\tilde{r}} d_k s_k', & \tilde{r} < \bar{r}, \\ \max(r_x, q) \sum_{k=1}^{\tilde{r}} s_k + \sum_{k=1}^{\tilde{r}} \sum_{l \neq k}^{\tilde{r}} \frac{d_k^2(s_k - s_l)}{d_k^2 - d_l^2} + \sum_{k=1}^{\tilde{r}} d_k s_k', & \tilde{r} = \bar{r}, \end{cases}$$

where for simplicity we write $s_k = s_k(d_k, \lambda)$ *and* $s_k' = \partial s_k(d_k, \lambda)/\partial d_k$.

The explicit formulae presented in the above theorems facilitate further exploration of the behaviour and properties of the degrees of freedom. For example, consider the unbiased estimator for reduced-rank regression in Theorem 3. It is always true that

$$\hat{\mathrm{df}}(r) \geqslant \max(r_x, q)r + \sum_{k=1}^{r} \sum_{l=r+1}^{\bar{r}} \frac{d_k^2 + 0}{d_k^2 - 0} = (r_x + q - r)r \quad (r = 1, \ldots, \bar{r}). \tag{12}$$

This suggests that the proposed estimator is always greater than the naive estimator, i.e., the number of free parameters $(r_x + q - r)r$. Similar to the lasso in univariate regression problems (Tibshirani, 1996; Zou et al., 2007), reduced-rank estimation can be viewed as a latent factor selection procedure, in which we both construct and search over as many as $\bar{r}$ latent linear factors. Therefore, the increments in the degrees of freedom as shown in (12) can be interpreted as the price we have to pay for performing this latent factor selection. For the general methods considered in Theorem 4, this inequality no longer holds, due to the shrinkage effects induced by the weights $0 \leqslant s_k \leqslant 1$. The reduction in the degrees of freedom due to singular value shrinkage can offset the price paid for searching over the set of latent variables. Therefore, similar to lasso, adaptive singular-value penalization can provide effective control over the model complexity (Tibshirani & Taylor, 2011).

Although the unbiased and naive estimators are quite different, there are some interesting connections between them. For instance, they are close to each other when evaluated at the true underlying rank, especially when the signal is strong relative to the noise level. This phenomenon has also been noted in empirical studies. Suppose that the true model rank is $\mathrm{rank}(B) = r^*$. Intuitively, the $\bar{r} - r^*$ smallest singular values from least squares may be close to zero and are not comparable to the r^* largest ones; using the approximation $d_k \approx 0$ $(k = r^* + 1, \ldots, \bar{r})$, we obtain $\hat{\mathrm{df}}(r^*) \approx (r_x + q - r^*)r^*$. A more rigorous argument can be made based on classical large-sample settings where p and q are fixed and $n \to \infty$; under standard assumptions, consistency of least-squares estimation can readily be established (Reinsel & Velu, 1998). Using techniques such as perturbation expansion of matrices (Izenman, 1975), the consistency of $\hat{Y}$ implies the consistency of the estimated singular values; that is, the first r^* estimated singular values converge to their nonzero true counterparts, while the rest converge to zero in probability. It follows that

$$\hat{\mathrm{df}}(r^*) \to (r_x + q - r^*)r^*, \quad n \to \infty,$$

in probability. An immediate implication is that for each $r = 1, \ldots, \bar{r}$, if we assume the true model to be of rank r, then in a classical asymptotic sense the number of free parameters, $(r_x + q - r)r$, is the correct degrees of freedom. This clearly relates to the error degrees of freedom of the classical asymptotic χ^2 statistic from the likelihood ratio test of $H_0 : \mathrm{rank}(B) = r$ for each

$r = 1, \ldots, \bar{r}$ (Izenman, 1975). If p and q are allowed to diverge with the sample size n, these classical asymptotic results would break down, as the convergence of nonzero singular values fails (Bai & Silverstein, 2009). In contrast, the unbiasedness property of our proposed exact estimator is non-asymptotic and hence valid for any given (p, q, n). Recently, non-asymptotic prediction error bounds have been developed for reduced-rank estimation methods, and the minimax convergence rate is found to coincide with the number of free parameters (Bunea et al., 2011; Rohde & Tsybakov, 2011). These results provide further justification of the proposed unbiased estimator and reveal the limitations, underlying assumptions and asymptotic nature of the naive estimator.

The derived formulae also reveal some interesting behaviour of rank reduction. In essence, reduced-rank methods distinguish the signal from the noise by examining the estimated singular values from least-squares estimation: the large singular values are more likely to represent the signals, while the small singular values mostly correspond to noise. By rank reduction, we aim to recover the signals that exceed a certain noise level. Suppose that d_k and d_{k+1} are close for some $k = 1, \ldots, \bar{r} - 1$. It can be argued that the true model rank is unlikely to be k, because the $(k + 1)$th and the kth layers are hardly distinguishable. Indeed, this is reflected in the degrees of freedom: for $r = k$, the formula includes a term $(d_k + d_{k+1})/(d_k - d_{k+1})$, which can be excessively large; on the other hand, there is no such term for $r = k + 1$. Consequently, the unbiased estimator of the degrees of freedom may not increase monotonically as the rank r increases, in contrast to the naive estimator. In the above scenario, the estimates for $r = k$ can even be larger than for $r = k + 1$. This automatically reduces the chance of k being selected as the final rank.

6. Simulation studies

6·1. *Unbiasedness*

In this simulation, our aim is to show that the degrees-of-freedom estimator defined via Theorem 3 is unbiased and can be significantly larger than the naive estimator, which simply counts the number of free parameters. Here unbiasedness is defined over the error distribution, and we treat X as a fixed design matrix. We conduct the study for both low and high dimensions. The parameters are as follows.

Setting I: $n = 100$, $p = 20$, $q = 12$, $r_0 = 6$.
Setting II: $n = 40$, $p = 80$, $q = 50$, $r_0 = 10$.

Here r_0 denotes the true rank of B. Let Σ denote the covariance matrix of the predictor variables X, and set $\Sigma_{jj'} = 0{\cdot}3^{|j-j'|}$. The rows of the predictor matrix are generated independently from $N_p(0, \Sigma)$. To control the singular structure of B through the covariance of signals XB, $B^{\mathrm{T}}\Sigma B$, we take the left singular vectors of B to be the same as the eigenvectors of Σ; the right singular vectors of B are generated by orthogonalizing a random standard normal matrix. The difference between successive nonzero singular values of B is fixed at 2. The error matrix is generated from independent and identical standard normal distributions. We replicate the process 200 times with a fixed design matrix. We compare the proposed exact method with the data-perturbation technique of Ye (1998) and the Monte Carlo estimator of the true degrees of freedom computed from (3). For the data-perturbation method, we consider 50 perturbations of the response matrix for each replication to estimate the partial derivatives numerically. We chose $0{\cdot}1\sigma$ for the perturbation size, where σ is the error standard deviation. Ideally we would expect the proposed exact estimator to be fairly close, on average, to the data-perturbation and Monte Carlo estimators. We compare the estimators against the naive degrees-of-freedom estimate, namely $\mathrm{df}_n(r) = (r_x + q - r)r$, which corresponds to the number of free parameters in a $p \times q$ matrix of rank r; this does not depend on the data.

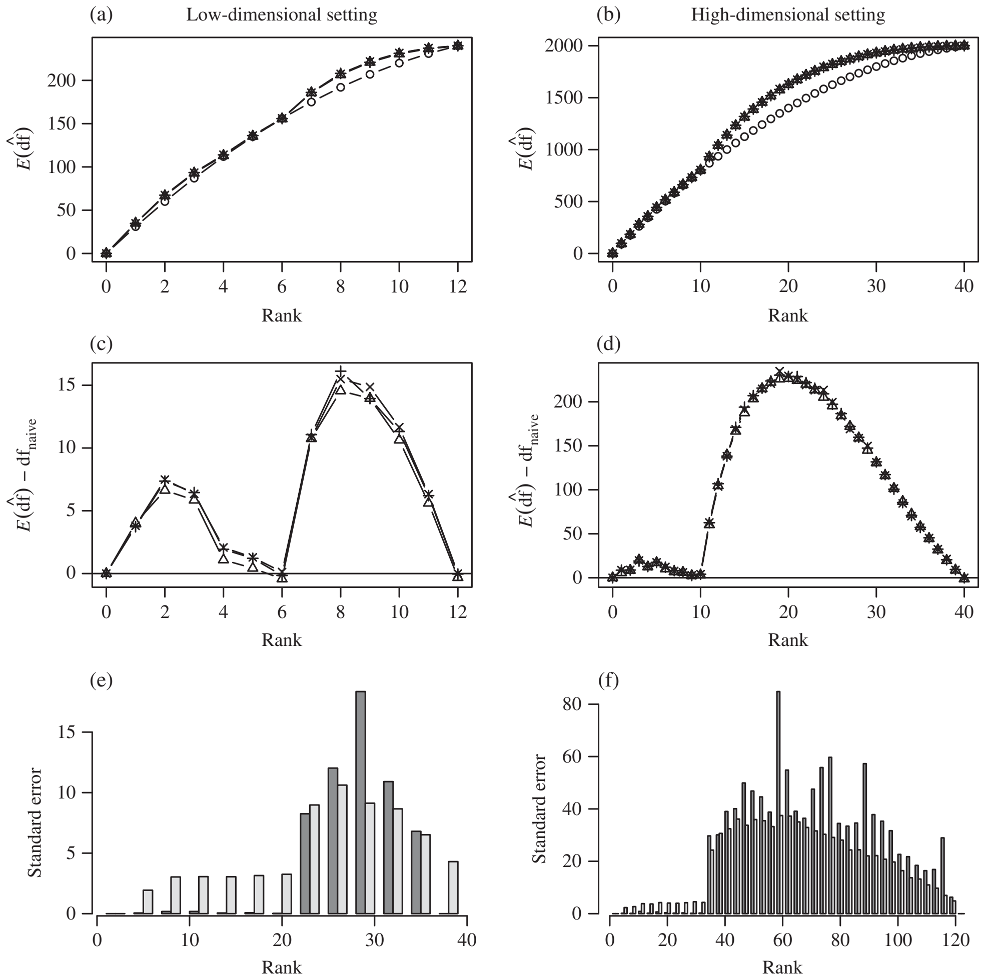

Fig. 1. Simulation results for the unbiased estimator of degrees of freedom in the low-dimensional setting (panels (a), (c) and (e)) and in the high-dimensional setting (panels (b), (d) and (f)). In panels (a) and (b), the average over 200 replications of the estimated degrees of freedom is plotted against the rank, for the exact estimator (asterisks), data-perturbation estimator (crosses), Monte Carlo estimator (triangles) and naive estimator (circles). Panels (c) and (d) plot the difference between the estimated degrees of freedom and the naive degrees of freedom for the same estimators. Panels (e) and (f) show the standard errors of the estimated degrees of freedom for the exact estimator (dark grey bars) and the perturbation estimator (light grey bars).

In Fig. 1(a) and (b) we see that for both high-dimensional and low-dimensional settings, the proposed exact estimator, the data-perturbation estimator and the Monte Carlo estimator yield nearly identical results. Further, as shown in Fig. 1(c) and (d), these estimates are significantly higher than the naive estimate; the difference is especially large once we go above the correct rank. This supports our theoretical intuition that the exact estimators seem to match the naive estimator very closely at the true rank. From Fig. 1(e) and (f) we can get a sense of the variability of the estimation procedures. The standard error for the proposed exact method is orders of magnitude smaller than that of data perturbation below the true rank; but once we go above the true rank, the standard error of the exact estimator becomes drastically greater. This arises from the fact that once we go above the true rank, the singular values of $\hat{Y}$ basically correspond to noise, and can be very close to each other. Hence, slight perturbations of the data may lead to different singular

Table 1. *Comparison of the prediction performance of different model selection criteria*

Setting	SNR	Measure	GCV[e]	GCV[n]	BIC[e]	BIC[n]
Low-dimensional	SNR ≈ 1	Est	1·56 (0·4)	1·80 (0·8)	1·56 (0·4)	1·76 (0·7)
		Pred	11·95 (2·2)	12·97 (3·4)	11·95 (2·2)	12·80 (2·2)
		Rank	3·01 (0·1)	3·18 (0·4)	3·01 (0·1)	3·12 (0·4)
	SNR $\approx 0{\cdot}5$	Est	6·00 (2·7)	7·47 (3·4)	5·92 (2·7)	7·31 (3·4)
		Pred	50·64 (10·8)	54·31 (10·8)	50·27 (10·8)	53·88 (10·7)
		Rank	2·41 (0·6)	2·86 (0·6)	2·34 (0·6)	2·77 (0·6)
High-dimensional	SNR ≈ 1	Est	3·25 (0·5)	3·30 (0·5)	3·10 (0·5)	3·30 (0·5)
		Pred	22·89 (1·5)	28·28 (4·3)	22·15 (1·4)	27·32 (4·1)
		Rank	4·84 (0·4)	5·30 (0·5)	4·75 (0·4)	5·22 (0·4)
	SNR $\approx 0{\cdot}5$	Est	3·77 (0·5)	4·00 (0·6)	3·61 (0·5)	3·90 (0·6)
		Pred	78·48 (6·2)	89·93 (17·4)	76·85 (6·1)	87·83 (17·2)
		Rank	4·00 (0·3)	4·46 (0·6)	3·92 (0·4)	4·34 (0·5)

[n], naive degrees-of-freedom estimator; [e], exact degrees-of-freedom estimator.

directions being selected, implying higher variability in model complexity. This phenomenon has also been noted by Ye (1998); that is, if one is trying to fit pure error components, the degrees of freedom tends to be high and unstable. This sudden change of standard deviation could be used as a tool to identify the underlying true rank. Presence of a strong changepoint in the variance profile would most certainly suggest a model of lower rank.

6·2. *Prediction performance*

Degrees-of-freedom estimates are commonly used in various model selection criteria. In this subsection we show that for reduced-rank regression, we can gain in prediction accuracy by using the exact degrees-of-freedom estimator instead of the naive one in a model selection criterion. We report the error metrics with respect to two model selection criteria, namely the generalized crossvalidation criterion (Golub et al., 1979) and the Bayesian information criterion (Schwarz, 1978). In the context of multivariate regression these are defined as follows:

$$\textsc{gcv}(r) = \frac{nq\,\|Y - \hat{Y}(r)\|_{\mathrm{F}}^2}{\{nq - \mathrm{df}(r)\}^2}, \quad \textsc{bic}(r) = \log\left\{\frac{1}{nq}\|Y - \hat{Y}(r)\|_{\mathrm{F}}^2\right\} + \frac{\log(nq)}{nq}\,\mathrm{df}(r).$$

We select the model that minimizes the model selection criterion over $1 \leqslant r \leqslant \min(n, r_x, q)$. Once again, we choose two settings for a comprehensive comparison.

Low-dimensional setting: $n = 50$, $p = 12$, $q = 10$, $r_0 = 3$.
High-dimensional setting: $n = 40$, $p = 80$, $q = 50$, $r_0 = 5$.

For each setting we consider two different levels of error variance, $\sigma^2 = 1$ and $\sigma^2 = 4$. This allows us to control the signal-to-noise ratio, $\textsc{snr} = d_{r_*}(XB)/d_1(\epsilon)$; the numerator is the smallest nonzero singular value of the signal matrix, a measure of the signal strength, and the denominator is the largest singular value of the error matrix, which is a measure of the noise strength (Bunea et al., 2011). Correlation among prediction variables is kept at a moderate level of 0·5. The data-generation scheme remains the same as in § 6·1. We report the estimation error $\mathrm{Est} = 100\|B - \hat{B}\|_{\mathrm{F}}^2/(pq)$, the prediction error $\mathrm{Pred} = 100\|XB - X\hat{B}\|_{\mathrm{F}}^2/(nq)$ and the selected rank. Table 1 summarizes the results. We report the averages over 100 replications, with standard errors in parentheses.

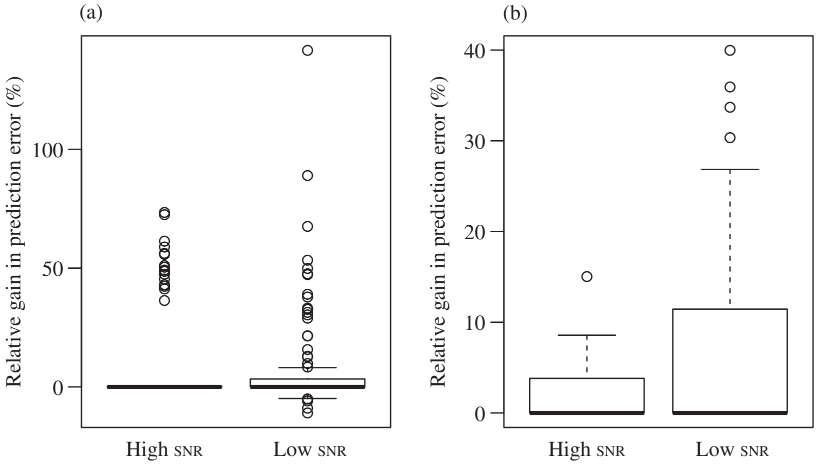

Fig. 2. Boxplots of the relative gain in prediction error obtained with the exact degrees-of-freedom estimator as compared to the naive estimator, in (a) the low-dimensional setting and (b) the high-dimensional setting. In each panel the two columns correspond to high and low levels of signal-to-noise ratio, SNR.

Using the proposed exact degrees-of-freedom estimator leads to better performance in terms of both GCV and BIC. The proposed estimator results in lower average estimation error and prediction error in all the settings. The relative gain is larger for the prediction error. Similar results, not shown here, were obtained at other levels of correlation. For the low-dimensional setting where an estimator for σ^2 is available, we also studied the performance of Mallows' C_P criterion; once again the results were very close to those reported and are thus omitted. We observe that BIC and GCV give nearly identical results, but BIC provides a slightly higher degree of regularization, especially in high-dimensional settings. When the signal-to-noise ratio is moderate to high, the naive degrees-of-freedom estimator tends to overestimate the rank, leading to inflated error measures. On the other hand, in low-signal settings, often the smallest nonzero singular values have very little explanatory power, and therefore selecting a lower-rank model enables us to do better in terms of prediction accuracy due to the bias-variance trade-off. As the exact degrees-of-freedom estimator usually gives higher values than the naive estimator, it penalizes the model complexity more strictly and leads to the selection of simpler models that yield better prediction. The difference between the mean prediction errors reported in Table 1 are within the ranges of the standard errors shown in parentheses. In order to demonstrate the superior prediction performance of the exact degrees-of-freedom estimator, we conduct a comparison on a per-dataset basis in terms of relative gain in accuracy. For each dataset, the pairwise relative gain is defined as

$$100 \times \frac{\text{Pred[n]} - \text{Pred[e]}}{\text{Pred[e]}}\%,$$

where Pred[e] and Pred[n] denote the prediction errors when using the exact and naive degrees-of-freedom estimators, respectively. Since the results for GCV and BIC were very similar, we plot only the results for GCV. As can be seen in Fig. 2, the boxplots tend to stay above zero, indicating that the exact degrees-of-freedom estimator outperforms the naive estimator consistently. Also, the relative gain is larger in the high-dimensional scenario.

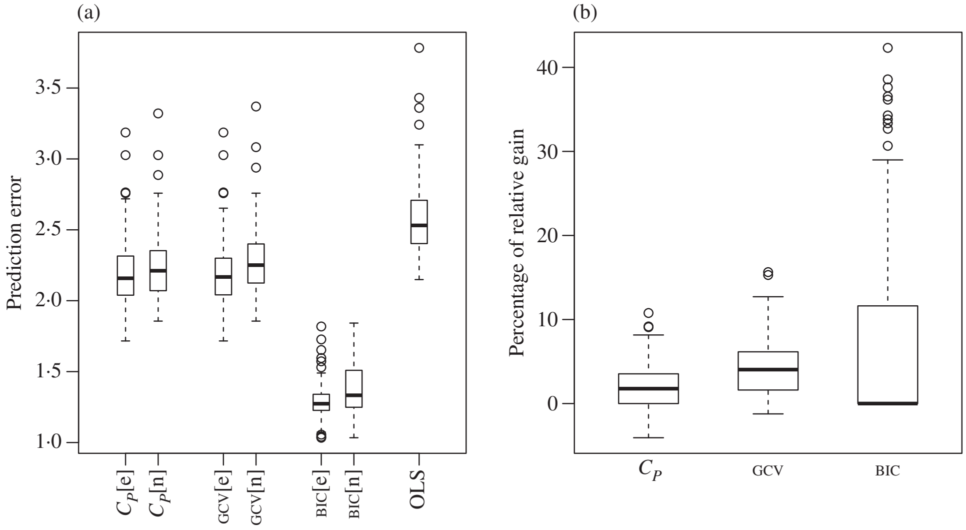

Fig. 3. Boxplots of prediction performance in the *Arabidopsis thaliana* data application: (a) mean squared prediction error of each method over 100 random splits for three model selection criteria, where [e] denotes use of the proposed exact degrees-of-freedom estimator and [n] denotes use of the naive degrees-of-freedom estimator in computing the model selection criteria; (b) relative gain in prediction error for each model selection method from using the exact degrees-of-freedom estimator over the naive degrees-of-freedom estimator.

7. Analysis of Arabidopsis thaliana data

In this section, we use the proposed method to select a reduced-rank model in the genetic association study of Wille et al. (2004). The goal of this microarray experiment was to understand the regulatory control mechanisms in the isoprenoid gene network of the plant *Arabidopsis thaliana*, more commonly known as thale cress or mouse-ear cress. Isoprenoids have many important biochemical functions in plants. To monitor the gene expression levels, 118 GeneChip microarray experiments were carried out. The predictors consist of 39 genes from two isoprenoid biosynthesis pathways, MVA and MEP, and the responses consist of the expression levels of 795 genes from 56 metabolic pathways, many of which are downstream of the two pathways considered as predictors. Thus some of the responses are expected to show significant association with the predictor genes. We select two downstream pathways, the carotenoid and phytosterol pathways, as our responses. It has already been shown experimentally that the carotenoid pathway is strongly linked to the MEP pathway, whereas the phytosterol pathway is significantly related to the MVA pathway; see Wille et al. (2004) and the references therein for a detailed discussion. Finally, we have 118 observations on $p = 39$ predictors and $q = 36$ responses, all logarithmically transformed to reduce the skewness. We also standardized the responses to make them comparable.

We split the dataset randomly into training and test sets of equal size. We fitted the model using the training samples and then used it to predict on the test set. The performance measure is the mean squared prediction error

$$\frac{2}{nq}\|Y_{\text{test}} - \hat{Y}_{\text{test}}\|_{\text{F}}^{2}.$$

The entire process was repeated 100 times based on random splits. We used Mallows' C_P, GCV and BIC with the exact degrees of freedom and the naive degrees of freedom to select the optimal rank.

Table 2. *Prediction accuracy and rank selection performance of the competing methods on the Arabidopsis thaliana data*

	C_P[e]	C_P[n]	GCV[e]	GCV[n]	BIC[e]	BIC[n]	OLS
Mean(Pred. err.)	2·20	2·24	2·19	2·282	1·30	1·39	2·59
Std(Pred. err.)	0·3	0·25	0·25	0·246	0·1	0·2	0·3
Mean(Est. rank)	8·76	9·71	8·68	10·52	1·09	1·48	–
Std(Est. rank)	1·2	0·8	1·3	1·0	0·4	0·8	–

Pred. err., prediction error; Est. rank, estimated rank; Std, standard error; [n], naive degrees-of-freedom estimator; [e], exact degrees-of-freedom estimator; OLS, ordinary least squares.

The mean squared prediction errors for each method are summarized in Fig. 3. For all three model selection criteria considered, the exact unbiased estimator outperforms the naive estimator in prediction accuracy. The relative gain is almost always positive, as shown by Fig. 3(b). Among the three model selection criteria, BIC appears to be the clear winner in terms of prediction error, by virtue of its selecting a very parsimonious model; see Table 2.

8. Concluding remarks

We have proposed an exact unbiased estimator of the degrees of freedom for a general class of reduced-rank estimators for the multivariate linear regression model. The proposed estimator can be computed explicitly, leading to a model selection procedure that is more efficient than the computationally expensive crossvalidation or data-perturbation methods. The closed form also gives us insight into the connection between the exact and naive degrees-of-freedom estimators. The proposed method assumes no conditions on the dimensions of the problem or the rank of the design matrix, and is suitable for application to high-dimensional problems where $p, q > n$. The methods developed can be extended to other related estimation procedures that employ regularization of the singular values, such as reduced-rank ridge regression (Mukherjee & Zhu, 2011).

There are several possible directions for future research. Although we have demonstrated numerically that using the unbiased degrees-of-freedom estimator improves model selection, it remains difficult to theoretically justify and quantify this performance gain in the finite-sample setting. We have mainly considered reduced-rank estimators which share the same set of singular vectors with the least-squares solution. It would be interesting to extend the results to other reduced-rank methods, such as nuclear-norm-penalized regression (Yuan et al., 2007). Since reduced-rank estimation can be more effective when combined with sparse estimation, e.g., selecting latent factors of a sparse subset of original variables, it would be very interesting to extend the method to sparse and low-rank models (Zou et al., 2007; Bunea et al., 2012; Chen et al., 2012). Another challenging problem is to extend the solutions to the case of general correlated errors, which is often encountered in practice. This is much more difficult, as $\hat{B}$ does not admit a closed-form solution for a general unknown error covariance matrix Σ. One possible approach would be to employ an iterative generalized least-squares-type algorithm to converge to the exact degrees of freedom. This also leads to the more general problem of investigating the degrees of freedom in rank-constrained generalized linear models (Yee & Hastie, 2003; Li & Chan, 2007; She, 2013). Finally, as reduced-rank methods are commonly used in the analysis of multiple time series, the proposed approach can be extended to such settings, for example to reduced-rank models with multiple sets of regressors (Velu, 1991) and co-integration (Anderson, 2002a).

Acknowledgement

The authors thank the editor, the associate editor, and two reviewers for their thoughtful comments and helpful suggestions. Chen was partially supported by the U.S. National Institutes of Health. Wang was partially supported by the National Cancer Institute. Zhu was supported by the U.S. National Institutes of Health and National Science Foundation.

Appendix

Proof of Equation (10). Upon using the Moore–Penrose inverse to simplify the ordinary least-squares solution, we get $\hat{Y} = XQS^{-1}H$ and $\hat{Y}(r) = XQS^{-1}H(r)$ for $r = 1, \ldots, \bar{r}$. Using the trace identity $\text{tr}(AB) = \text{tr}(BA)$, the equality $\text{vec}(ABC) = (C^{\mathrm{T}} \otimes A)\text{vec}(B)$ and the chain rule of differentiation, we obtain

$$\begin{aligned}\hat{\text{df}}(r) &= \text{tr}\left[\frac{\partial\,\text{vec}\{\hat{Y}(r)\}}{\partial\,\text{vec}(Y)}\right]\\ &= \text{tr}\left(\left(I_q \otimes XQS^{-1}\right)\left[\frac{\partial\,\text{vec}\{H(r)\}}{\partial\,\text{vec}(Y)}\right]\right)\\ &= \text{tr}\left(\left(I_q \otimes XQS^{-1}\right)\left[\frac{\partial\,\text{vec}\{H(r)\}}{\partial\,\text{vec}(H)}\right]\left\{\frac{\partial\,\text{vec}(H)}{\partial\,\text{vec}(Y)}\right\}\right)\\ &= \text{tr}\left(\left(I_q \otimes XQS^{-1}\right)\left[\frac{\partial\,\text{vec}\{H(r)\}}{\partial\,\text{vec}(H)}\right]\left(I_q \otimes S^{-1}Q^{\mathrm{T}}X^{\mathrm{T}}\right)\right)\\ &= \text{tr}\left[\frac{\partial\,\text{vec}\{H(r)\}}{\partial\,\text{vec}(H)}\right].\end{aligned}$$ □

Proof of Theorem 1. The proof is based mainly on results of Magnus & Neudecker (1998) concerning the derivatives of a generalized eigensystem. We have assumed $r_x \geqslant q$; the same results can be stated for H^{T} when $r_x \leqslant q$. Let $A = H^{\mathrm{T}}H$, and let (d^2, v) denote an eigenvalue-eigenvector pair for A. Suppose that A is twice continuously differentiable at θ, e.g., $\theta = h_{ij}$ for any $i = 1, \ldots, r_x$ and $j = 1, \ldots, q$. Then the eigenvalues and eigenvectors are also differentiable at θ. As $Av = d^2 v$, it follows that

$$\frac{\partial A}{\partial\theta}v + A\frac{\partial v}{\partial\theta} = d^2\frac{\partial v}{\partial\theta} + \frac{\partial d^2}{\partial\theta}v,$$

and this gives

$$(A - d^2 I)\frac{\partial v}{\partial\theta} = -\left(\frac{\partial A}{\partial\theta} - \frac{\partial d^2}{\partial\theta}I\right)v. \tag{A1}$$

Premultiplying both sides of (A1) by v^{T} gives

$$v^{\mathrm{T}}(A - d^2 I)\frac{\partial v}{\partial\theta} = -v^{\mathrm{T}}\frac{\partial A}{\partial\theta}v + \frac{\partial d^2}{\partial\theta}.$$

It is obvious that the left-hand side equals zero; it then follows that

$$\frac{\partial d}{\partial\theta} = \frac{1}{2d}v^{\mathrm{T}}\frac{\partial A}{\partial\theta}v. \tag{A2}$$

Define $(A - d^2 I)^- = V(D^2 - d^2 I)^- V^{\mathrm{T}}$, where $(\cdot)^-$ denotes the Moore–Penrose inverse. Then $(A - d^2 I)^-(A - d^2 I) = I - vv^{\mathrm{T}}$ and $(A - d^2 I)^- v = 0$. Premultiplying both sides of (A1) by $(A - d^2 I)^-$ gives

$$(I - vv^{\mathrm{T}})\frac{\partial v}{\partial\theta} = -(A - d^2 I)^-\frac{\partial A}{\partial\theta}v.$$

As $v^{\mathrm{T}}v = 1$, we have that $v^{\mathrm{T}}(\partial v/\partial\theta) = 0$, and so

$$\frac{\partial v}{\partial\theta} = -(A - d^2 I)^{-}\frac{\partial A}{\partial\theta}v. \tag{A3}$$

Define $Z^{(ij)} = \partial H/\partial h_{ij}$, a $r_x \times q$ matrix of zeros with only its (i, j)th entry equal to 1. For any $\theta = h_{ij}$,

$$\frac{\partial A}{\partial h_{ij}} = H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H. \tag{A4}$$

The proof is completed by combining (A2), (A3) and (A4). □

Proof of Theorem 3. For simplicity and without loss of generality, we assume $r_x \geqslant q$. When $r_x \leqslant q$, one can repeat the same proof using H^{T}. When $r = q$, the statement $\widehat{\mathrm{df}}(q) = r_x q$ holds trivially. So in the following we assume $r < q$. Consider $\partial H^{(r)}/\partial h_{ij}$ for any $1 \leqslant i \leqslant r_x$ and $1 \leqslant j \leqslant q$. Because $H^{(r)} = H\sum_{k=1}^{r} v_k v_k^{\mathrm{T}}$, by the chain rule we have that

$$\begin{aligned}\frac{\partial H^{(r)}}{\partial h_{ij}} &= \frac{\partial H}{\partial h_{ij}}\sum_{k=1}^{r} v_k v_k^{\mathrm{T}} + H\sum_{k=1}^{r}\frac{\partial v_k}{\partial h_{ij}}v_k^{\mathrm{T}} + H\sum_{k=1}^{r} v_k\frac{\partial v_k^{\mathrm{T}}}{\partial h_{ij}}\\ &= Z^{(ij)}V^{(r)}V^{(r)\mathrm{T}} - H\sum_{k=1}^{r}\left\{(H^{\mathrm{T}}H - d_k^2 I)^{-}(H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H)v_k v_k^{\mathrm{T}}\right\}\\ &\quad - H\sum_{k=1}^{r}\left\{v_k v_k^{\mathrm{T}}(H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H)(H^{\mathrm{T}}H - d_k^2 I)^{-}\right\}.\end{aligned} \tag{A5}$$

Consider the first term on the right-hand side of (A5). Its (i, j)th entry equals $\sum_{k=1}^{r} v_{jk}^2$, so its contribution to the degrees of freedom (10) is

$$\sum_{i=1}^{r_x}\sum_{j=1}^{q}\sum_{k=1}^{r} v_{jk}^2 = r_x r, \tag{A6}$$

because $\sum_{j=1}^{q} v_{jk}^2 = 1$. We know that

$$(H^{\mathrm{T}}H - d_k^2 I)^{-} = \sum_{l \neq k}^{q}\frac{1}{d_l^2 - d_k^2}v_l v_l^{\mathrm{T}},$$

and we also have

$$H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H = \begin{pmatrix} & & h_{i1} & & \\ & & \vdots & & \\ h_{i1} & \cdots & 2h_{ij} & \cdots & h_{iq}\\ & & \vdots & & \\ & & h_{iq} & & \end{pmatrix}.$$

Now consider the second term on the right-hand side of (A5). After some algebra, its (i, j)th entry can be written as $u_i^{\mathrm{T}}Da^{(ij)}$, where $a^{(ij)} \in \mathbb{R}^q$ with

$$a_k^{(ij)} = -\sum_{l \neq k}^{r}\frac{1}{d_k^2 - d_l^2}(v_{jk}v_{jl}h_i^{\mathrm{T}}v_l + v_{jl}^2 h_i^{\mathrm{T}}v_k) \quad (k = 1, \ldots, q).$$

Similarly, the (i, j)th entry of the third term on the right-hand side of (A5) is $u_i^{\mathrm{T}}Db^{(ij)}$, where $b^{(ij)} \in \mathbb{R}^q$ with

$$b_k^{(ij)} = -\sum_{l \neq k}^{q}\frac{1}{d_l^2 - d_k^2}(v_{jk}v_{jl}h_i^{\mathrm{T}}v_l + v_{jl}^2 h_i^{\mathrm{T}}v_k) \quad (k = 1, \ldots, r)$$

and $b_k^{(ij)} = 0$ for $k = r+1, \ldots, q$ whenever $r < q$. Consider the second and third terms together. Since

$$a_k^{(ij)} + b_k^{(ij)} = \begin{cases} \displaystyle\sum_{l=r+1}^{q} \frac{1}{d_k^2 - d_l^2} (v_{jk} v_{jl} h_i^{\mathrm{T}} v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) & (k = 1, \ldots, r), \\ \displaystyle\sum_{l=1}^{r} \frac{1}{d_l^2 - d_k^2} (v_{jk} v_{jl} h_i^{\mathrm{T}} v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) & (k = r+1, \ldots, q), \end{cases}$$

it follows that the contribution from the second and third terms equals

$$\begin{aligned} &\sum_{i=1}^{r_x} \sum_{j=1}^{q} \left\{ \sum_{k=1}^{r} u_{ik} d_k \sum_{l=r+1}^{q} \frac{1}{d_k^2 - d_l^2} (v_{jk} v_{jl} h_i^{\mathrm{T}} v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) \right\} \\ &\quad + \sum_{i=1}^{r_x} \sum_{j=1}^{q} \left\{ \sum_{k=r+1}^{q} u_{ik} d_k \sum_{l=1}^{r} \frac{1}{d_l^2 - d_k^2} (v_{jk} v_{jl} h_i^{\mathrm{T}} v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) \right\} \\ &= \sum_{i=1}^{r_x} \left\{ \sum_{k=1}^{r} u_{ik} d_k \sum_{l=r+1}^{q} \frac{1}{d_k^2 - d_l^2} \sum_{j=1}^{q} (v_{jk} v_{jl} h_i^{\mathrm{T}} v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) \right\} \\ &\quad + \sum_{i=1}^{r_x} \left\{ \sum_{k=r+1}^{q} u_{ik} d_k \sum_{l=1}^{r} \frac{1}{d_l^2 - d_k^2} \sum_{j=1}^{q} (v_{jk} v_{jl} h_i^{\mathrm{T}} v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) \right\} \\ &= \sum_{i=1}^{r_x} \left\{ \sum_{k=1}^{r} \sum_{l=r+1}^{q} \frac{d_k}{d_k^2 - d_l^2} u_{ik} (h_i^{\mathrm{T}} v_k) + \sum_{k=r+1}^{q} \sum_{l=1}^{r} \frac{d_k}{d_l^2 - d_k^2} u_{ik} (h_i^{\mathrm{T}} v_k) \right\} \\ &= \sum_{i=1}^{r_x} \left\{ \sum_{k=1}^{r} \sum_{l=r+1}^{q} \frac{d_k}{d_k^2 - d_l^2} u_{ik} (h_i^{\mathrm{T}} v_k) + \sum_{k=1}^{r} \sum_{l=r+1}^{q} \frac{d_l}{d_k^2 - d_l^2} u_{il} (h_i^{\mathrm{T}} v_l) \right\} \\ &= \sum_{k=1}^{r} \sum_{l=r+1}^{q} \left(\frac{d_k}{d_k^2 - d_l^2} u_k^{\mathrm{T}} H v_k + \frac{d_l}{d_k^2 - d_l^2} u_l^{\mathrm{T}} H v_l \right) \\ &= \sum_{k=1}^{r} \sum_{l=r+1}^{q} \frac{d_k^2 + d_l^2}{d_k^2 - d_l^2}. \end{aligned}$$

Upon combining this with (A6), the proof is completed. □

Proof of Theorem 4. Again, we assume $r_x \geqslant q$. When $r_x \leqslant q$, one can repeat the same proof using H^{T}. Recall that $\tilde{H}(\lambda) = U\tilde{D}(\lambda)V^{\mathrm{T}}$. Consider $\partial \tilde{H}(\lambda)/\partial h_{ij}$ for any fixed $\lambda > 0$, $1 \leqslant i \leqslant r_x$ and $1 \leqslant j \leqslant q$. Let $\tilde{r} = \tilde{r}(\lambda) = \max\{k : s_k > 0\}$. Because $\tilde{H}(\lambda) = H \sum_{k=1}^{\tilde{r}} s_k v_k v_k^{\mathrm{T}}$, by the chain rule we have that

$$\begin{aligned} \frac{\partial \tilde{H}(\lambda)}{\partial h_{ij}} &= \frac{\partial H}{\partial h_{ij}} \sum_{k=1}^{\tilde{r}} s_k v_k v_k^{\mathrm{T}} + H \sum_{k=1}^{\tilde{r}} s_k \frac{\partial v_k}{\partial h_{ij}} v_k^{\mathrm{T}} + H \sum_{k=1}^{\tilde{r}} s_k v_k \frac{\partial v_k^{\mathrm{T}}}{\partial h_{ij}} + H \sum_{k=1}^{\tilde{r}} \frac{\partial s_k}{\partial h_{ij}} v_k v_k^{\mathrm{T}} \\ &= Z^{(ij)} V^{(\tilde{r})} D^{(\tilde{r})-1} \tilde{D}^{(\tilde{r})} V^{(\tilde{r})\mathrm{T}} - H \sum_{k=1}^{\tilde{r}} \left\{ s_k (H^{\mathrm{T}} H - d_k^2 I)^{-} (H^{\mathrm{T}} Z^{(ij)} + Z^{(ij)\mathrm{T}} H) v_k v_k^{\mathrm{T}} \right\} \\ &\quad - H \sum_{k=1}^{\tilde{r}} \left\{ s_k v_k v_k^{\mathrm{T}} (H^{\mathrm{T}} Z^{(ij)} + Z^{(ij)\mathrm{T}} H)(H^{\mathrm{T}} H - d_k^2 I)^{-} \right\} \\ &\quad + H \sum_{k=1}^{\tilde{r}} \left[s_k' \left\{ \frac{1}{2 d_k} v_k^{\mathrm{T}} (H^{\mathrm{T}} Z^{(ij)} + Z^{(ij)\mathrm{T}} H) v_k \right\} v_k v_k^{\mathrm{T}} \right], \end{aligned} \qquad (\mathrm{A}7)$$

where $s'_k = \partial s_k/\partial d_k$. It can be shown that the (i, j)th entry of the first term on the right-hand side of (A7) equals $\sum_{k=1}^{\tilde{r}} s_k v_{jk}^2$, so its contribution to the degrees of freedom (10) is

$$\sum_{i=1}^{r_x}\sum_{j=1}^{q}\sum_{k=1}^{\tilde{r}} s_k v_{jk}^2 = r_x \sum_{k=1}^{\tilde{r}} s_k, \tag{A8}$$

because $\sum_{j=1}^{q} v_{jk}^2 = 1$. Similar to the proof of Theorem 3, the (i, j)th entry of the second and third terms on the right-hand side of (A7) can be shown to be $u_i^{\mathrm{T}} D(\tilde{a}^{(ij)} + \tilde{b}^{(ij)})$, where $\tilde{a}^{(ij)} \in \mathbb{R}^q$ and $\tilde{b}^{(ij)} \in \mathbb{R}^q$ with

$$\tilde{a}_k^{(ij)} + \tilde{b}_k^{(ij)} = \begin{cases} \displaystyle\sum_{l \neq k}^{q} \frac{s_k - s_l}{d_k^2 - d_l^2}(v_{jk}v_{jl}h_i^{\mathrm{T}}v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) & (k = 1, \ldots, \tilde{r}), \\ \displaystyle\sum_{l=1}^{\tilde{r}} \frac{s_l}{d_l^2 - d_k^2}(v_{jk}v_{jl}h_i^{\mathrm{T}}v_l + v_{jl}^2 h_i^{\mathrm{T}} v_k) & (k = \tilde{r} + 1, \ldots, q). \end{cases}$$

After some algebra, one can show that the contribution of the second and third terms to the degrees of freedom equals

$$\begin{aligned} &\sum_{i=1}^{r_x}\sum_{j=1}^{q}\left\{\sum_{k=1}^{\tilde{r}} u_{ik}d_k \sum_{l \neq k}^{q} \frac{s_k - s_l}{d_k^2 - d_l^2}(v_{jk}v_{jl}h_i^{\mathrm{T}}v_l + v_{jl}^2 h_i^{\mathrm{T}}v_k)\right\} \\ &\quad + \sum_{i=1}^{r_x}\sum_{j=1}^{q}\left\{\sum_{k=\tilde{r}+1}^{q} u_{ik}d_k \sum_{l=1}^{\tilde{r}} \frac{s_l}{d_l^2 - d_k^2}(v_{jk}v_{jl}h_i^{\mathrm{T}}v_l + v_{jl}^2 h_i^{\mathrm{T}}v_k)\right\} \\ &\quad = \sum_{k=1}^{\tilde{r}}\sum_{s=\tilde{r}+1}^{q}\left\{\frac{d_k^2(s_k - s_l) + d_l^2 s_k}{d_k^2 - d_l^2}\right\} + \sum_{k=1}^{\tilde{r}}\sum_{l \neq k}^{\tilde{r}}\left\{\frac{d_k^2(s_k - s_l)}{d_k^2 - d_l^2}\right\}. \end{aligned} \tag{A9}$$

Consider the fourth term on the right-hand side of (A7). Note that $v_k^{\mathrm{T}}(H^{\mathrm{T}}Z^{(ij)} + Z^{(ij)\mathrm{T}}H)v_k = 2v_{jk}(v_k^{\mathrm{T}}h_i)$. The (i, j)th entry of the fourth term is $\sum_{k=1}^{\tilde{r}} s'_k u_{ik} v_{jk}^2 (v_k^{\mathrm{T}} h_i)$, so the contribution of the fourth term to the degrees of freedom is

$$\sum_{i=1}^{r_x}\sum_{j=1}^{q}\sum_{k=1}^{\tilde{r}} s'_k u_{ik} v_{jk}^2 (v_k^{\mathrm{T}}h_i) = \sum_{i=1}^{r_x}\sum_{k=1}^{\tilde{r}} s'_k u_{ik}(v_k^{\mathrm{T}}h_i) = \sum_{k=1}^{\tilde{r}} s'_k \sum_{i=1}^{r_x} u_{ik}h_i^{\mathrm{T}}v_k = \sum_{k=1}^{\tilde{r}} d_k s'_k.$$

Upon combining this with (A8) and (A9), the proof is completed. □

Proof of Theorem 2. We start with a few definitions and facts from algebraic geometry and matrix analysis.

Definition A1. *An algebraic variety over* $\mathbb{R}^k$ *(or* $\mathbb{C}^k$*) is defined as the set of points satisfying a system of polynomial equations* $\{f_\ell(x_1, x_2, \ldots, x_k) = 0 : \ell \in \mathcal{I}\}$.

Here each $f_\ell(\cdot)$ is a polynomial function of its arguments and $\mathcal{I}$ denotes an index set. If at least one of the $f_\ell(\cdot)$ is not identically zero, it is called a proper subvariety. A proper subvariety must be of dimension less than k and therefore has Lebesgue measure zero in $\mathbb{R}^k$ (Allman et al., 2009). For a more detailed discussion, we refer the reader to Hartshorne (1977) or Cox et al. (2007).

Proposition A1 (Laub, 2004). *Any square symmetric matrix* $M \in \mathbb{R}^{k\times k}$ *has at least one repeated eigenvalue if and only if* $\operatorname{rank}(M \otimes I_k - I_k \otimes M) < (k^2 - k)$.

Define

$$\mathcal{S}_1 = \{A \in \mathbb{R}^{r_x \times q} : A \text{ has at least one singular value that is } 0\},$$
$$\mathcal{S}_2 = \{A \in \mathbb{R}^{r_x \times q} : A \text{ has at least one repeated singular value}\}.$$

Note that $\mathcal{S}^{\mathrm{c}} = \mathcal{S}_1 \cup \mathcal{S}_2$, so it is enough to show that $\mu(S_1) = 0$ and $\mu(S_2) = 0$. By Definition A1 and the discussion above, it suffices to show that $\mathcal{S}_1$ and $\mathcal{S}_2$ are proper subvarieties of $\mathbb{R}^{r_x \times q}$. The set $\mathcal{S}_1$ can be rewritten as

$$\mathcal{S}_1 = \{A \in \mathbb{R}_{r_x \times q} : \det(A^{\mathrm{T}}A) = 0\},$$

where $\det(\cdot)$ denotes the determinant of a square matrix. Now, $\det(A^{\mathrm{T}}A)$ is a nontrivial polynomial in the entries of A, and hence $\mathcal{S}_1$ is a proper subvariety and has Lebesgue measure zero. For $\mathcal{S}_2$, observe that if $A \in \mathbb{R}^{p \times q}$ has at least one repeated singular value, then $A^{\mathrm{T}}A \in \mathbb{R}^{p \times q}$ must have at least one repeated eigenvalue. Therefore, in view of Proposition A1, $\mathcal{S}_2$ can be reformulated as

$$\mathcal{S}_2 = \{A \in \mathbb{R}_{r_x \times q} : \mathrm{rank}(A^{\mathrm{T}}A \otimes I_q - I_q \otimes A^{\mathrm{T}}A) < (q^2 - q)\}.$$

This is an algebraic variety, since it can be expressed as the solution to all minors of order at least $q^2 - q$ being equal to zero, which are all polynomial equations in the entries of A. Thus we have shown that $\mu(\mathcal{S}_1 \cup \mathcal{S}_2) = 0$. □

References

Allman, E. S., Matias, C. & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37**, 3099–132.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327–51.

Anderson, T. W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *Ann. Statist.* **27**, 1141–54.

Anderson, T. W. (2002a). Reduced rank regression in cointegrated models. *J. Economet.* **106**, 203–16.

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

Bai, Z. D. & Silverstein, J. W. (2009). *Spectral Analysis of Large Dimensional Random Matrices*. New York: Springer, 2nd ed.

Breiman, L. & Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *J. R. Statist. Soc.* B **59**, 3–37.

Bunea, F., She, Y. & Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39**, 1282–309.

Bunea, F., She, Y. & Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.* **40**, 2359–88.

Chen, K., Chan, K. S. & Stenseth, N. R. (2012). Reduced-rank stochastic regression with a sparse singular value decomposition. *J. R. Statist. Soc.* B **74**, 203–21.

Chen, K., Dong, H. & Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika* **100**, 901–20.

Chen, L. & Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Am. Statist. Assoc.* **107**, 1533–45.

Cox, D. A., Little, J. & O'Shea, D. (2007). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. New York: Springer, 3rd ed.

Davies, P. T. & Tso, M. K.-S. (1982). Procedures for reduced-rank regression. *Appl. Statist.* **31**, 244–55.

Donoho, D. & Johnstone, I. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Assoc.* **90**, 1200–24.

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1**, 211–8.

Efron, B., Burman, P., Denby, L., Landwehr, J. M., Mallows, C. L., Shen, X., Huang, H.-C., Ye, J. & Zhang, C. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with Comments, Rejoinder). *J. Am. Statist. Assoc.* **99**, 619–42.

Golub, G., Heath, M. & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–23.

Hartshorne, R. (1977). *Algebraic Geometry*. New York: Springer.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. London: Chapman & Hall.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hocking, R. R. & Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics* **9**, 531–40.

Hotelling, H. (1935). The most predictable criterion. *J. Educ. Psychol.* **26**, 139–42.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Mult. Statist.* **5**, 248–64.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*. New York: Springer.

Laub, A. J. (2004). *Matrix Analysis for Scientists and Engineers*. Philadelphia: Society for Industrial and Applied Mathematics.

Li, M. C. & Chan, K. S. (2007). Multivariate reduced-rank nonlinear time series modeling. *Statist. Sinica* **17**, 139–59.

Li, Y. & Zhu, J. (2008). L_1-norm quantile regression. *J. Comp. Graph. Statist.* **17**, 163–85.

Lu, Z., Monteiro, R. & Yuan, M. (2012). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Math. Programming* **131**, 163–94.

Magnus, J. R. & Neudecker, H. (1998). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661–75.

Massy, W. F. (1965). Principal components regression with exploratory statistical research. *J. Am. Statist. Assoc.* **60**, 234–46.

Meyer, M. & Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28**, 1083–104.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.* **26**, 394–5.

Mukherjee, A. & Zhu, J. (2011). Reduced rank ridge regression and its kernel extensions. *Statist. Anal. Data Mining* **4**, 612–22.

Neghaban, S. & Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39**, 1069–97.

O'Neil, K. (2005). Critical points of the singular value decomposition. *SIAM J. Matrix Anal. Appl.* **27**, 459–73.

Penrose, R. (1955). A generalized inverse for matrices. *Proc. Camb. Phil. Soc.* **51**, 406–13.

Rao, C. R. (1978). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Proc. 5th Int. Symp. Mult. Anal.*, P. R. Krishnaiah, ed. Amsterdam: North-Holland, pp. 3–22.

Reinsel, G. C. & Velu, R. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer.

Rohde, A. & Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39**, 887–930.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.

She, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Statist.* **3**, 384–415.

She, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statist. Infer.* **6**, 197–209.

Shen, X. & Ye, J. (2002). Adaptive model selection. *J. Am. Statist. Assoc.* **97**, 210–21.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1135–51.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Tibshirani, R. J. & Taylor, J. (2011). Degrees of freedom in lasso problems. *Ann. Statist.* **40**, 1198–232.

Tsukuma, H. (2008). Admissibility and minimaxity of Bayes estimators for a normal mean matrix. *J. Mult. Anal.* **99**, 2251–64.

Velu, R. (1991). Reduced rank models with two sets of regressors. *Appl. Statist.* **40**, 159–70.

Wille, A., Zimmermann, P., Vranova, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W. & Bühlmann, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.* **5**, 1–13.

Witten, D. M., Tibshirani, R. J. & Hastie, T. J. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**, 515–34.

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*, J. Gani, ed. London: Academic Press.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Am. Statist. Assoc.* **93**, 120–31.

Yee, T. & Hastie, T. J. (2003). Reduced rank vector generalized linear models. *Statist. Mod.* **3**, 15–41.

Yuan, M., Ekici, A., Lu, Z. & Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc.* B **69**, 329–46.

Zou, H., Hastie, T. J. & Tibshirani, R. J. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173–92.

[*Received April* 2013. *Revised June* 2014]