Biometrika (2017), **104**, 1, *pp.* 97–110
Printed in Great Britain
doi: 10.1093/biomet/asw059
Advance Access publication 3 January 2017

Selective factor extraction in high dimensions

By YIYUAN SHE

Department of Statistics, Florida State University, 117 N Woodward Ave., Tallahassee, Florida 32306, U.S.A.

yshe@stat.fsu.edu

Summary

This paper studies simultaneous feature selection and extraction in supervised and unsupervised learning. We propose and investigate selective reduced rank regression for constructing optimal explanatory factors from a parsimonious subset of input features. The proposed estimators enjoy sharp oracle inequalities, and with a predictive information criterion for model selection, they adapt to unknown sparsity by controlling both rank and row support of the coefficient matrix. A class of algorithms is developed that can accommodate various convex and nonconvex sparsity-inducing penalties, and can be used for rank-constrained variable screening in high-dimensional multivariate data. The paper also showcases applications in macroeconomics and computer vision to demonstrate how low-dimensional data structures can be effectively captured by joint variable selection and projection.

Some key words: Information criterion; Nonconvex optimization; Oracle inequality; Principal component analysis; Reduced rank regression; Variable screening.

1. Introduction

Modern statistical applications may involve many variables. Principal component analysis (Hotelling, 1933) offers a popular means of dimension reduction, and reduced rank regression extends it to supervised learning (Anderson, 1951) by solving the problem $\min_{B\in\mathbb{R}^{p\times m}} \|Y - XB\|_{\mathrm{F}}^2$ subject to $r(B) \leqslant r$, where $Y \in \mathbb{R}^{n\times m}$ and $X \in \mathbb{R}^{n\times p}$ are response and predictor matrices, $r(B)$ denotes the rank of B, and $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. Reduced rank regression provides a low-dimensional projection space to view and analyse multivariate data, and finds widespread applications in machine learning, econometrics, and finance (Reinsel & Velu, 1998; Izenman, 2008). In fact, once an estimate B of rank r is obtained, we can write $B = B_1 B_2^{\mathrm{T}}$ for $B_1 \in \mathbb{R}^{p\times r}$, $B_2 \in \mathbb{R}^{m\times r}$. This suggests that r factors can be constructed by XB_1 from p predictors to explain all response variables. The number of factors required in real applications is often much smaller than the number of input x-variables. Unfortunately, the loading matrix B_1 obtained from reduced rank regression typically involves all predictors. In high-dimensional data analysis, factors constructed from a small subset of variables are much more interpretable; we call this selective factor extraction. Correspondingly, the coefficient matrix is desired to have both low rank and row-wise sparsity. To capture the two types of structural parsimony simultaneously, joint regularization must be applied, which adds nontrivial difficulties in the theoretical analysis and numerical computation of the associated estimators, but leads to reduced errors compared with rank reduction or variable selection alone.

In the unsupervised setting when $X = I$, selective factor extraction is closely related to sparse principal component analysis. See, for example, Zou et al. (2006), Shen & Huang (2008), Witten et al. (2009), Johnstone & Lu (2009) and Ma (2013). Most of these algorithms seek sparse loading vectors separately, and progress sequentially. The loading matrix obtained may lack optimality and contain too many variables. To ensure dimension reduction even when constructing a number of factors, we will formulate the problem as a whole and pursue joint sparsity across all loading vectors. This turns out be particularly helpful in rank-constrained variable screening.

There has been less work on simultaneous variable selection and rank reduction in the supervised setting; see Bunea et al. (2012), Chen & Huang (2012), Chen et al. (2012), Ma et al. (2014) and a recent report by Ma et al. (arXiv:1403.1922). Many theoretical and computational questions remain open. Our main contributions are three-fold. First, we are able to provide a unified treatment of various penalties in the reduced rank model, and successfully build sharper oracle inequalities than those in the literature (Bickel et al., 2009; Lounici et al., 2011; Candès & Plan, 2011). Our results indicate that for joint variable selection and rank reduction, the error rates and parameter choices previously obtained are suboptimal. Second, we develop a computational framework with guaranteed convergence, where any thresholding rule can be applied. The algorithms adapt to reduced rank variable screening in very high dimensions. Third, we come up with a new information criterion for parameter tuning. To the best of our knowledge, this is the first sound criterion with minimax optimality for selecting among sparse and/or rank-deficient models.

In the rest of the paper, the following notation and symbols will be used. Given a matrix $A = (\alpha_1, \ldots, \alpha_p)^{\mathrm{T}} \in \mathbb{R}^{p\times m}$, $\|A\|_{\mathrm{F}}$ and $\|A\|_2$ denote its Frobenius norm and spectral norm, respectively. We define the $(2,1)$-norm of A by $\|A\|_{2,1} = \sum_{j=1}^{p} \|\alpha_j\|_2$, and use $\|A\|_{2,0} = \sum_{j=1}^{p} 1_{\|\alpha_j\|\neq 0}$ to characterize the number of nonzero rows in A. The standard vectorization of A is denoted by vec (A). We use $A(\mathcal{J}, \mathcal{I})$ to denote a submatrix of A with rows and columns indexed by $\mathcal{J}$ and $\mathcal{I}$, respectively, and occasionally abbreviate $A(\mathcal{J}, \cdot)$ to $A_{\mathcal{J}}$. The set of column-orthogonal matrices of size $m \times r$ is denoted by $\mathbb{O}^{m\times r} = \{V \in \mathbb{R}^{m\times r} : V^{\mathrm{T}}V = I\}$. Finally, C and c are used to denote constants which are not necessarily the same at each occurrence.

2. SIMULTANEOUS RANK REDUCTION AND VARIABLE SELECTION

2·1. *Selective reduced rank regression*

Picking only pertinent dimensions is the key to enhancing interpretability of factors in high dimensions. In a multifactor model, power to select requires eliminating nuisance variables from the construction of factors. To state a general framework, we assume that a response matrix $Y \in \mathbb{R}^{n\times m}$ is available, in addition to a predictor matrix $X \in \mathbb{R}^{n\times p}$, both centred columnwise. Let B denote the coefficient matrix $B = (b_1, \ldots, b_p)^{\mathrm{T}} = (b_{j,k})$. To provide concurrent rank reduction and feature selection, a possible optimization criterion is $\min_{B\in\mathbb{R}^{p\times m}} \|Y - XB\|_{\mathrm{F}}^2 + \lambda_1^2 r(B) + \lambda_2^2 \|B\|_{2,0}$. But the penalized form does not seem to enjoy low errors in either theory or practice. We propose the following form of rank-constrained variable selection

$$\min_{B\in\mathbb{R}^{p\times m}} \frac{1}{2}\|Y - XB\|_{\mathrm{F}}^2 + \sum_{j=1}^{p} P(\|b_j\|_2; \lambda) \quad \text{subject to} \quad r(B) \leqslant r, \tag{1}$$

where P is a sparsity-promoting penalty, possibly nonconvex. We call (1) selective reduced-rank regression. Imposing elementwise sparsity on B, though valid as a regularization approach, does

not seem to have much meaning in applications. We will introduce a different sparse reduced rank regression in (11) by sparsifying a component of B.

There is a variety of choices for the penalty function. The popular group ℓ_1-norm function, $\lambda\|B\|_{2,1}$ (Yuan & Lin, 2006), leads to the rank-constrained group lasso, although the group ℓ_0 penalty, $(\lambda^2/2)\|B\|_{2,0}$, is arguably better at promoting sparsity (Bunea et al., 2012; Chen & Huang, 2012). Group versions of the nonconvex penalties proposed by Fan & Li (2001), Zhang (2010a), and Zhang (2010b) can also be applied.

Solving the selective reduced rank regression problem helps uncover factors that reduce model complexity. Given a selective reduced rank regression estimate $\hat{B}$, we can use its column space to make a new model matrix $Z = XUD$, where U, D and V are obtained from the singular value decomposition $\hat{B} = UDV^{\mathrm{T}}$. The new design has r columns and involves only a small subset of the x-variables. This is called Type I factor extraction. An alternative is to decompose $X\hat{B}$. Concretely, let $\hat{B}^{\mathrm{T}}X^{\mathrm{T}}X\hat{B} = VDV^{\mathrm{T}}$ be the spectral decomposition of $\hat{B}^{\mathrm{T}}X^{\mathrm{T}}X\hat{B}$. Then $Z = X\hat{B}V$ provides r factors, called Type II extraction or post-decorrelation, since the z-variables are uncorrelated with each other. QR decomposition can be used for efficiency reasons in either case. The two types of factor extraction are not equivalent in general, but coincide when $\hat{B}$ is the solution to reduced rank regression. Because $r \ll p$, a more sophisticated model can be built on the factors with relative ease.

2·2. *Sharp-rate oracle inequalities*

We show some non-asymptotic oracle inequalities to reveal the theoretical benefits of selective reduced rank regression. For clarity, we use the group ℓ_0 and group ℓ_1 penalties to exemplify the error rate. For $B = (b_1, \ldots, b_p)^{\mathrm{T}}$, define $\mathcal{J}(B) = \{j : b_j \neq 0\}$ and $J(B) = |\mathcal{J}(B)| = \|B\|_{2,0}$.

THEOREM 1. *Let $Y = XB^* + \mathcal{E}$, with all entries of $\mathcal{E}$ independent and identically distributed as $N(0, \sigma^2)$.*

(i) *Let $\hat{B}$ be a selective reduced rank regression estimator that minimizes $\|Y - XB\|_{\mathrm{F}}^2 + \lambda^2\|B\|_{2,0}$ subject to $r(B) \leqslant r$. Then, under $\lambda = A\sigma(r + \log p)^{1/2}$ where A is a large enough constant, the following oracle inequality holds for any $B \in \mathbb{R}^{p\times m}$ with $r(B) \leqslant r$:*

$$E(\|X\hat{B} - XB^*\|_{\mathrm{F}}^2) \lesssim \|XB - XB^*\|_{\mathrm{F}}^2 + \lambda^2 J(B) + (m - r)r\sigma^2 + \sigma^2. \tag{2}$$

Here, $\lesssim$ means that the inequality holds up to a multiplicative constant.

(ii) *In the ℓ_1 case, let $\hat{B} = \arg\min_{B:r(B)\leqslant r} \|Y - XB\|_{\mathrm{F}}^2/(2\|X\|_2) + \lambda\|B\|_{2,1}$ where λ is as in* (i). *Then $E(\|X\hat{B} - XB^*\|_{\mathrm{F}}^2) \lesssim \|XB - XB^*\|_{\mathrm{F}}^2 + K^2\lambda^2 J(B) + (m-r)r\sigma^2 + \sigma^2$ holds for any $B \in \mathbb{R}^{p\times m}$ with $r(B) \leqslant r$, provided that X satisfies*

$$(1 + \vartheta)\|X\|_2\|\Delta_{\mathcal{J}}\|_{2,1} \leqslant K|\mathcal{J}|^{1/2}\|X\Delta\|_{\mathrm{F}} + \|X\|_2\|\Delta_{\mathcal{J}^{\mathrm{c}}}\|_{2,1}, \quad \Delta \in \mathbb{R}^{p\times m} \tag{3}$$

where $\mathcal{J} = \mathcal{J}(B)$, $K \geqslant 0$, and ϑ is a positive constant.

The proof given in the Supplementary Material can deal with various penalties in a universal way. For example, the oracle inequality (2) applies to any $P(\cdot;\lambda)$ that takes λ as the threshold and satisfies $P_H(\theta;\lambda) \leqslant P(\theta;\lambda) \leqslant C\lambda^2$, where $P_H(\theta;\lambda) = (-\theta^2/2 + \lambda|\theta|)1_{|\theta|<\lambda} + (\lambda^2/2)1_{|\theta|\geqslant\lambda}$. Examples include the smoothly clipped absolute deviation penalty (Fan & Li, 2001), the minimax concave penalty (Zhang, 2010a) and the capped ℓ_1 penalty (Zhang, 2010b). Similarly, the result in part (ii) of Theorem 1 holds for any subadditive penalty that is sandwiched between $P_H(\theta;\lambda)$ and $P_1(\theta;\lambda) = \lambda|\theta|$. The ℓ_p penalties $P(\theta;\lambda) = (2-2p)^{1-p}(2-p)^{p-2}\lambda^{2-p}|\theta|^p$ where $0 < p < 1$

are particular instances. Moreover, condition (3) is less demanding than some common regularity assumptions (van de Geer & Bühlmann, 2009; She, 2016), and we do not require $\|X\|_2$ to be bounded above by $Cn^{1/2}$.

Let $r^* = r(B^*)$ and $J^* = |\mathcal{J}(B^*)|$. According to (2), simply taking $r = r^*$ and $B = B^*$ so that the bias term $\|XB - XB^*\|_{\mathrm{F}}^2$ disappears, we get a prediction error bound of order

$$(J^* + m - r^*)r^* + J^* \log p, \tag{4}$$

omitting σ^2 and constant factors. The bias term makes the error bound applicable to coefficient matrices that are approximately row-sparse. In § 4·2, we will see that when it is difficult to provide a proper rank value, the predictive information criterion can be used to tune r to guarantee the same low error rate.

A comparison between Theorem 1 and some existing non-asymptotic results follows. Wei & Huang (2010) and Lounici et al. (2011) showed that for group lasso, the prediction error is of the order $J^*m + J^* \log p$. Since $(J^* + m - r^*)r^* + J^* \log p \lesssim J^*m + J^* \log p$, selective rank reduction is uniformly better, and the performance gain is dramatic for low-rank models. Bunea et al. (2012) obtained an error rate for the rank-constrained group lasso of $J^*r^* \log p + mr^*$. Their rate is, however, suboptimal: when r^* and J^* are comparable, their error bound is of the order $J^{*2} \log p + J^*m$, while (4) gives $J^* \log p + J^*m$. Bunea et al. (2011) also required a multivariate restricted eigenvalue assumption that is more restrictive than (3). Compared with low-rank matrix estimation (Recht et al., 2010; Bunea et al., 2011), which has an error rate of $mr + qr$ with $q = r(X)$, our result does not always show an improvement, because only large values of A are considered in Theorem 1 to secure selectivity. Practically there will be no performance loss, because selective reduced rank regression degenerates to reduced rank regression when $\lambda = 0$.

3. PARAMETER TUNING AND MODEL COMPARISON

3·1. *A predictive information criterion*

Selective reduced rank regression has two regularization parameters, λ and r, to control the row support and rank of the model. Conventional tuning methods are not satisfactory in our experience, and indeed they all lack theoretical support in the sparse and rank-deficient setting. We will propose a novel information criterion from the perspective of predictive learning (Hastie et al., 2009), namely, the best model should give the smallest prediction error among all candidate models. Unlike consistent variable selection or rank selection, such a principle does not require high signal-to-noise ratios to work.

To make our results more general, the noise matrix is assumed to have sub-Gaussian marginal tails in this section. A random variable ξ is sub-Gaussian if $\mathrm{pr}(|\xi| \geqslant t) \leqslant C \exp(-ct^2)$ for any $t > 0$ and some constants $C, c > 0$, and its scale is defined as $\sigma(\xi) = \inf\{\sigma > 0 : E\{\exp(\xi^2/\sigma^2)\} \leqslant 2\}$. Gaussian random variables and bounded random variables are particular instances. More generally, $\xi \in \mathbb{R}^p$ is a sub-Gaussian random vector with its scale bounded by σ if $\langle \xi, \alpha \rangle$ is sub-Gaussian and $\sigma(\langle \xi, \alpha \rangle) \leqslant \sigma \|\alpha\|_2$ for any $\alpha \in \mathbb{R}^p$.

The function proposed as the model complexity penalty is

$$P_o(B) = \sigma^2[\{q \wedge J(B) + m - r(B)\}r(B) + J(B) \log\{ep/J(B)\}],$$

where $q = r(X)$ and $q \wedge J(B) = \min\{q, J(B)\}$.

THEOREM 2. *Assume that the vectorized noise matrix, or* vec $(\mathcal{E})$*, is sub-Gaussian with mean zero and scale bounded by* σ*. Let* $\hat{B} \in \arg\min_B \frac{1}{2}\|Y - XB\|_{\mathrm{F}}^2 + AP_o(B)$*, where* A *is a constant. Then for all sufficiently large values of* A*,* $\hat{B}$ *satisfies the oracle inequality*

$$E\left[\max\{\|X\hat{B} - XB^*\|_{\mathrm{F}}^2, P_o(\hat{B})\}\right] \lesssim \inf_{B \in \mathbb{R}^{p\times m}} \{\|XB - XB^*\|_{\mathrm{F}}^2 + P_o(B)\} + \sigma^2. \tag{5}$$

Theorem 2 is a strong non-asymptotic result because the obtained error rate is uniformly better than those given by selection or rank reduction as mentioned in § 2·2. Indeed, we can show that P_o gives the minimax optimal error rate in this jointly sparse setting. Moreover, (5) holds under no restrictions on X or B^*, and its right-hand side takes the infimum over all reference signals $B \in \mathbb{R}^{p\times m}$.

The theorem gives rise to a model comparison criterion. By the same reasoning as in its proof, for any collection of random nonzero matrices $B_1, \ldots, B_l, \ldots$, if we choose the optimal one, B_o, by minimizing the predictive information criterion over all given matrices,

$$\|Y - XB\|_{\mathrm{F}}^2 + AP_o(B), \tag{6}$$

then B_o satisfies $E(\|XB_o - XB^*\|_{\mathrm{F}}^2) \leqslant C \inf_{l\geqslant 1} E\{\|XB_l - XB^*\|_{\mathrm{F}}^2 + P_o(B_l)\}$. Interestingly, P_o indicates that $J(B)\log\{ep/J(B)\}$, the inflation term due to selection, should be additive to the degrees-of-freedom term. This is legitimate for sub-Gaussian noise contamination, but to our knowledge new when compared with other information criteria that take the form of loss + $c(n,p) \times$ degrees-of-freedom. For example, the extended Bayesian information criterion (Chen & Chen, 2008), derived under $p = O(n^\kappa)$ with $\kappa > 0$ and some other regularity conditions, has a multiplicative factor $\log n + \log p$ on the degrees-of-freedom of the model. For single-response models with $m = 1$, P_o simplifies to $\sigma^2[q \wedge J(B) + J(B)\log\{ep/J(B)\}]$, which essentially corresponds to the risk inflation criterion (Foster & George, 1994) but is slightly finer. Our result applies to any n, p, m.

3·2. *Scale-free predictive information criterion*

The predictive information criterion contains a scale parameter σ. In sparse principal component analysis, one can substitute an estimate $\hat{\sigma}$ for the unknown σ, e.g., $\hat{\sigma}^2 = \mathrm{med}(\|x_j\|_2^2/n)$ (Johnstone & Lu, 2009). In supervised learning, however, estimating the scale parameter could be as hard as estimating the coefficients. We propose a scale-free form of predictive information criterion that can bypass σ. Again, no incoherence assumption is made for the predictor matrix.

THEOREM 3. *Let* $\mathcal{E}$ *have independent and identically distributed* $N(0,\sigma^2)$ *entries. Suppose that the true model is parsimonious in the sense that* $P_o(B^*)/\sigma^2 < mn/A_0$ *for some constant* $A_0 > 0$*. Consider the criterion*

$$\|Y - XB\|_{\mathrm{F}}^2/\{mn - AP_o(B)/\sigma^2\}, \tag{7}$$

where the constant A *satisfies* $0 < A < A_0$*. Then, for sufficiently large values of* A_0 *and* A*, any* $\hat{B}$ *that minimizes* (7) *subject to* $P_o(B)/\sigma^2 < mn/A$ *satisfies* $\|X\hat{B} - XB^*\|_{\mathrm{F}}^2 \lesssim P_o(B^*)$*, with probability at least* $1 - Cp^{-c} - C'\exp(-c'mn)$ *for some constants* $C, C', c, c' > 0$*.*

The real model complexity penalty is of the form $A_1 \times$ degrees-of-freedom $+ A_2 \times$ inflation, with constants A_1 and A_2 that can be determined by computer experiments. Experience shows that when σ^2 is known or can be well-estimated, the choice $A_1 = 2{\cdot}4$, $A_2 = 1{\cdot}8$ works well in (6), and we recommend $A_1 = 2$, $A_2 = 1{\cdot}8$ for the scale-free form (7).

4. COMPUTATION

4·1. *A computational framework*

To ensure that selective reduced rank regression can be applied, we must address some challenges in computation. First, the rank constraint makes problem (1) nonconvex and nonsmooth. Moreover, in view of Theorem 1, to relax the incoherence conditions required by ℓ_1-type penalties, nonconvex penalties may be of interest (Zhang, 2010a; Zhang & Zhang, 2012). Since different nonconvex penalty forms may lead to the same thresholding rule, we study thresholding-induced penalties.

DEFINITION 1 (Threshold function). *A threshold function is a real-valued function* $\Theta(s;\lambda)$ *defined for* $-\infty < s < \infty$ *with* $\lambda \geqslant 0$ *as the parameter such functions that* (i) $\Theta(-s;\lambda) = -\Theta(s;\lambda)$, (ii) $\Theta(s;\lambda) \leqslant \Theta(s';\lambda)$ *for* $s \leqslant s'$, (iii) $\lim_{s\to\infty}\Theta(s;\lambda) = \infty$, *and* (iv) $0 \leqslant \Theta(s;\lambda) \leqslant s$ *for* $0 \leqslant s < \infty$. *Moreover,* $\vec{\Theta}$ *is defined to be a multivariate function associated with* Θ *if for any vector* $a \in \mathbb{R}^m$, $\vec{\Theta}(a;\lambda) = a\Theta(\|a\|_2;\lambda)/\|a\|_2$ *for* $a \neq 0$ *and* 0 *otherwise. For any matrix* $A \in \mathbb{R}^{p\times m}$ *with* $A = (a_1\ \ldots\ a_p)^{\mathrm{T}}$, $\vec{\Theta}(A;\lambda) = \{\vec{\Theta}(a_1;\lambda)\ \ldots\ \vec{\Theta}(a_p;\lambda)\}^{\mathrm{T}}$.

Some thresholding functions, such as the hard-thresholding $\Theta(s;\lambda) = s1_{|s|\leqslant\lambda}$ or $s1_{|s|<\lambda}$, have discontinuities. To avoid ambiguity in the definition, when using such functions, we assume that the quantity to be thresholded does not correspond to a discontinuity point. Let us consider the following scaled version of problem (1):

$$\min_{B=(b_1,\ldots,b_p)^{\mathrm{T}}\in\mathbb{R}^{p\times m}} F(B;\lambda) = \frac{1}{2K}\|Y - XB\|_{\mathrm{F}}^2 + \sum_{j=1}^{p} P(\|b_j\|_2;\lambda) \quad \text{subject to } r(B) \leqslant r, \tag{8}$$

where P is associated with Θ through (9) and K is a large enough number to be specified in Theorem 4. To get rid of the low-rank constraint, we may write $B = SV^{\mathrm{T}}$, with $S = (s_1,\ldots,s_p)^{\mathrm{T}} \in \mathbb{R}^{p\times r}$ and $V \in \mathbb{O}^{m\times r} = \{V \in \mathbb{R}^{m\times r} : V^{\mathrm{T}}V = I\}$. The optimization is now with respect to V and s_j $(j = 1,\ldots,p)$. We abuse notation and use $F(s_1,\ldots,s_p,V;\lambda)$ to denote the objective function. Algorithm 1 is developed based on a block coordinate descent method, where the V-optimization is solved by Procrustes rotation and the s_j is obtained by iterative thresholding.

Algorithm 1. Selective reduced rank regression.

Data: $1 \leqslant r \leqslant p$; $\lambda \geqslant 0$; $S^{(0)} \in \mathbb{R}^{p\times r}$; Θ, thresholding rule; M_{inner}, maximum number of inner iterations; M_{outer}, maximum number of outer iterations.

1) $t \leftarrow 0$, $K \leftarrow \|X\|_2^2$

repeat

2) $t \leftarrow t + 1$

3) Let $W \leftarrow Y^{\mathrm{T}}XS^{(t-1)}$, and compute its reduced singular value decomposition $W = U_wD_wV_w^{\mathrm{T}}$

4) $V^{(t)} \leftarrow U_wV_w^{\mathrm{T}}$

5) Execute the subroutine below to update S

5a) $l \leftarrow 0$, $\tilde{S}^{(0)} \leftarrow S^{(t-1)}$

repeat

5b) $l \leftarrow l + 1$

5c) $\Xi^{(l,t)} \leftarrow X^{\mathrm{T}}YV^{(t-1)}/K + (I - X^{\mathrm{T}}X/K)\tilde{S}^{(l-1)}$

5d) $\tilde{S}^{(l)} \leftarrow \vec{\Theta}(\Xi^{(l,t)};\lambda)$

until $l \geqslant M_{\text{inner}}$ or $\|\tilde{S}^{(l)} - \tilde{S}^{(l-1)}\|$ is sufficiently small

6) $S^{(t)} \leftarrow \tilde{S}^{(l)}$
7) $B^{(t)} \leftarrow S^{(t)}(V^{(t)})^{\mathrm{T}}$
until $t \geqslant M_{\text{outer}}$ and $\|B^{(t)} - B^{(t-1)}\|$ is sufficiently small.
Output $\hat{B} = B^{(t)}$, $\hat{V} = V^{(t)}$.

We will show that, given any Θ, the algorithm is guaranteed to converge under a universal choice of K. For simplicity, in the following theorem we assume that $\vec{\Theta}(\cdot;\lambda)$ is continuous at any point in the closure of $\{\Xi^{(l,t)} : l \geqslant 1, t \geqslant 1\}$. The condition holds for all continuous thresholding rules. Practically used thresholding rules have few discontinuity points and such discontinuities rarely occur in real data analysis.

THEOREM 4. *Given an arbitrary thresholding function $\Theta(\cdot;\lambda)$, let P be an associated penalty satisfying*

$$P(\theta;\lambda) - P(0;\lambda) = \int_0^{|\theta|} [\sup\{s : \Theta(s;\lambda) \leqslant u\} - u]\,\mathrm{d}u + Q(\theta;\lambda), \tag{9}$$

for some Q satisfying $Q(\cdot,\lambda) \geqslant 0$ and $Q(\theta;\lambda) = 0$ if $\theta = \Theta(s;\lambda)$ for some $s \in \mathbb{R}$. Let $K \geqslant \|X\|_2^2$. Then given any starting point $S^{(0)} \in \mathbb{R}^{p\times r}$, $F(B^{(t)};\lambda)$ converges, $r(B^{(t)}) \leqslant r$, and

$$F(B^{(t)}) - F(B^{(t+1)}) \geqslant (1 - \|X\|_2^2/K)\|S^{(t)} - S^{(t+1)}\|_{\mathrm{F}}^2/2.$$

Furthermore, if $K > \|X\|_2^2$, then any accumulation point of $(s_1^{(t)}, \ldots, s_p^{(t)}, V^{(t)})$ is a coordinatewise minimum point of F and the function value converges monotonically to $F(s_1^*, \ldots, s_p^*, V^*)$ for some coordinatewise minimum point $(s_1^*, \ldots, s_p^*, V^*)$.

Equation (9) covers all aforementioned convex and nonconvex penalties; see She (2012) for more examples. For penalties with $Q(\cdot;\lambda) = 0$, Theorem 4 provides a stationary point guarantee. When Θ has discontinuities, Q can have infinitely many choices, which means that different penalties may be associated with the same thresholding function. For instance, define a hard-ridge thresholding rule

$$\Theta_{\mathrm{HR}}(s;\lambda,\eta) = \begin{cases} 0, & |s| < \lambda, \\ s/(1+\eta), & |s| \geqslant \lambda. \end{cases} \tag{10}$$

Then, with a nontrivial Q defined by $Q(\theta;\lambda,\eta) = 0{\cdot}5(1+\eta)(\lambda - |\theta|)^2 1_{0<|\theta|<\lambda}$, (9) gives an $\ell_0+\ell_2$ penalty $P(\theta;\lambda,\eta) = \eta\theta^2/2 + \lambda^2 1_{\theta\neq 0}/(2+2\eta)$, or $P(B;\lambda,\eta) = \eta\|B\|_{\mathrm{F}}^2/2 + \lambda^2\|B\|_{2,0}/(2+2\eta)$ in the context of (1). The Frobenius component in the hybrid penalty can shrink the coefficients to compensate for collinearity and large noise in large-p applications. Section 4·2 makes use of a constraint variant of the penalty for screening.

When we apply a componentwise Θ in place of $\vec{\Theta}$ in Step 5d, a result similar to Theorem 4 can be obtained for the objective function

$$\min_{S=(s_{j,k})\in\mathbb{R}^{p\times r}, V\in\mathbb{O}^{m\times r}} \|Y - XSV^{\mathrm{T}}\|_{\mathrm{F}}^2/(2K) + \sum_{j=1}^{p}\sum_{k=1}^{r} P(|s_{j,k}|;\lambda^e). \tag{11}$$

The sparsity is imposed on S rather than on the overall coefficient matrix SV^{T}. We call (11) sparse reduced rank regression. With $S = (\tilde{s}_1, \ldots, \tilde{s}_r)$ and $V = (v_1, \ldots, v_r)$, we see that XSV^{T} is a sum of r factors, $X\tilde{s}_1 v_1^{\mathrm{T}} + \cdots + X\tilde{s}_r v_r^{\mathrm{T}}$, and every $\tilde{s}_k$ is sparse ($k = 1, \ldots, r$).

Algorithm 1 is simple to implement and has low computational complexity. When $r > 1$, in addition to some elementary matrix multiplication and thresholding operations, a singular value decomposition is carried out on W, which, however, has only r columns. To initialize the algorithm, we can use the reduced rank regression estimate $(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}YV_rV_r^{\mathrm{T}}$ and set $S^{(0)} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}YV_r$, where $^{+}$ denotes the Moore–Penrose inverse and V_r is formed by the first r eigenvectors of $Y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y$. Other initialization schemes are possible; see Rousseeuw & Van Driessen (1999).

From Algorithm 1, or the proof of Theorem 4 in the Supplementary Material, the optimal V can be expressed in terms of S, i.e., $V_o(S) = U_wV_w^{\mathrm{T}} = \{(XS)^{\mathrm{T}}YY^{\mathrm{T}}XS\}^{-1/2}(XS)^{\mathrm{T}}Y$. Hence $\|Y - XSV_o^{\mathrm{T}}(S)\|_{\mathrm{F}}^2 = \|[I - XS\{S^{\mathrm{T}}X^{\mathrm{T}}YY^{\mathrm{T}}XS\}^{-1/2}(XS)^{\mathrm{T}}]Y\|_{\mathrm{F}}^2$ or $\|XS\|_{\mathrm{F}}^2 - 2\|Y^{\mathrm{T}}XS\|_{*} + \|Y\|_{\mathrm{F}}^2$, where $\|\cdot\|_{*}$ is the nuclear norm. This means that the loading matrix obtained from (8) or (11) depends on Y through YY^{T}.

Some recent theoretical studies (Berthet & Rigollet, 2013; Gao et al., 2017) show that computationally efficient algorithms, such as those with polynomial time complexity, may possess an intrinsic lower bound in statistical accuracy that is larger than the minimax error rate derived for most challenging problems. This seems to hold in our problem as well. We will not pursue this further in the current paper.

4·2. *Rank-constrained variable screening*

Statisticians are frequently confronted with challenges in large-scale computing, so variable screening has become a popular practice in high-dimensional data analysis. In multivariate problems, we are interested in rank-constrained variable screening, which can be achieved by the following form of selective reduced rank regression

$$\min_{B\in\mathbb{R}^{p\times m}} F(B) = \frac{1}{2K}\|Y - XB\|_{\mathrm{F}}^2 + \frac{\eta}{2}\|B\|_{\mathrm{F}}^2 \quad \text{subject to } \|B\|_{2,0} \leqslant d, r(B) \leqslant r. \tag{12}$$

Similar to the rank constraint, which limits the number of factors, the cardinality constraint, rather than a penalty, enables one to directly control the number of predictors selected for factor construction. The upper bound d can be loose for the purpose of screening, provided it is not too small.

We show below how to use a quantile version of the hard-ridge thresholding (10) to solve such problems. Given $1 \leqslant d \leqslant p$, $\eta \geqslant 0$, for any $s = (s_1, \ldots, s_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\Theta^{\#}(s; d, \eta)$ is defined to be a vector $t = (t_1, \ldots, t_p)^{\mathrm{T}} \in \mathbb{R}^p$ satisfying $t_{(j)} = s_{(j)}/(1 + \eta)$ if $1 \leqslant j \leqslant d$ and 0 otherwise. Here, $s_{(1)}, \ldots, s_{(p)}$ are the order statistics of $s_1, \ldots, s_p$, i.e., $|s_{(1)}| \geqslant \cdots \geqslant |s_{(p)}|$, and $t_{(1)}, \ldots, t_{(p)}$ are defined similarly. In the case of ties, a random tie-breaking rule is used. The multivariate quantile thresholding function $\vec{\Theta}^{\#}(S; d, \eta)$ for any $S = (s_1, \ldots, s_p)^{\mathrm{T}} \in \mathbb{R}^{p\times r}$ is defined as a $p \times r$ matrix $T = (t_1, \ldots, t_p)^{\mathrm{T}}$ with $t_j = s_j/(1 + \eta)$ if $\|s_j\|_2$ is among the d largest elements in $\{\|s_j\|_2 : 1 \leqslant s \leqslant p\}$ and 0 otherwise. Now we modify Step 5d of Algorithm 1 to $\tilde{S}^{(l)} \leftarrow \vec{\Theta}^{\#}(\Xi^{(l,t)}; d, \eta)$, while all other steps remain unchanged. The resulting algorithm for rank-constrained screening always converges.

THEOREM 5. *Assume $K \geqslant \|X\|_2^2$. Then, given any $S^{(0)} \in \mathbb{R}^{p\times r}$, $F(B^{(t)})$ is nonincreasing and satisfies $F(B^{(t)}) - F(B^{(t+1)}) \geqslant (1 - \|X\|_2^2/K)\|S^{(t)} - S^{(t+1)}\|_{\mathrm{F}}^2/2$, and $B^{(t)}$ obeys the constraints $\|B^{(t)}\|_{2,0} \leqslant d$ and $r(B^{(t)}) \leqslant r$ for any $t \geqslant 1$.*

To gain some intuition, let us set $S^{(0)} = 0$. Then, at the first iteration, $W = 0$, $V^{(1)} = I$, and the quantile thresholding picks d features according to the marginal statistics $X^{\mathrm{T}}Y$, which

amounts to sure independence screening (Fan & Lv, 2008). Our algorithm iterates further to lessen the greediness of independence screening. To accelerate the computation, we recommend progressive screening in the iterative process. Concretely, we use a sequence $Q(t)$ that decreases from p to d, e.g., $Q(t) = 2p/\{1 + \exp(\alpha t)\}$ with $\alpha = 0{\cdot}01$ and $0 \leqslant t \leqslant (1/\alpha)\log(2p/d - 1)$, and perform $\tilde{S}^{(l)} \leftarrow \vec{\Theta}^{\#}(\Xi^{(l,t)}; Q(t), \eta)$ in Step 5d; after obtaining $B^{(t)}$ in Step 7, the following data-squeezing operations are carried out: $\mathcal{J} \leftarrow \{j : S^{(t)}(j, 1:r) \neq 0\}$, $S^{(t)} \leftarrow S^{(t)}(\mathcal{J}, 1:r)$, $X \leftarrow X(1:n, \mathcal{J})$. An attractive feature of the implementation is that as the cycles progress, the problem size drops quickly and the computational load can be significantly reduced.

For the sparse reduced rank regression with an ℓ_0-constraint,

$$\min_{S \in \mathbb{R}^{p\times r}, V \in \mathbb{O}^{m\times r}} \frac{1}{2K}\|Y - XSV^{\mathrm{T}}\|_{\mathrm{F}}^2 + \frac{\eta}{2}\|S\|_{\mathrm{F}}^2 \quad \text{subject to } \|S\|_0 \leqslant d^e, \tag{13}$$

similar algorithms can be developed based on iterative quantile thresholding. In big data applications, a good idea is to combine (12) with (13), because calling the rank-constrained screening algorithm in an earlier stage can reduce the dimensionality from p to d. In this hybrid scheme, d satisfies $d \leqslant d^e \leqslant dr$, and $d = d^e$ gives a conservative screening choice.

5. Unsupervised selective and sparse principal component analyses

This section studies selective factor construction in principal component analysis. We assume that only one data matrix $X \in \mathbb{R}^{n\times p}$ is available and it has been column-centred. Principal component analysis can be interpreted as finding a low-rank matrix B to approximate the observed data. Similar to § 4·1, we write $B = VS^{\mathrm{T}}$ with $S = (s_{j,k}) = (s_1, \ldots, s_p)^{\mathrm{T}} \in \mathbb{R}^{p\times r}$ and $V \in \mathbb{O}^{n\times r}$. The selective principal component analysis problem is defined as

$$\min_{S \in \mathbb{R}^{p\times r}, V \in \mathbb{O}^{n\times r}} \frac{1}{2}\|X - VS^{\mathrm{T}}\|_{\mathrm{F}}^2 + \sum_{j=1}^{p} P(\|s_j\|_2; \lambda). \tag{14}$$

Obviously, (14) can be rephrased as a special case of selective reduced rank regression, by taking X^{T} as the response matrix and $I_{p\times p}$ as the design matrix. Likewise, adapting (11) to the unsupervised setting leads to the following criterion for sparse principal component analysis:

$$\min_{S \in \mathbb{R}^{p\times r}, V \in \mathbb{O}^{p\times r}} \frac{1}{2}\|X - VS^{\mathrm{T}}\|_{\mathrm{F}}^2 + \sum_{j=1}^{p}\sum_{k=1}^{r} P(|s_{j,k}|; \lambda^e). \tag{15}$$

Unsupervised versions of (12) and (13) can also be defined. Moreover, based on the discussions in § 4·1, all these criteria depend on X through $X^{\mathrm{T}}X$.

Problems (14) and (15) are perhaps less challenging than their supervised counterparts, but they still provide new insights into sparse principal component analysis. For example, (15) defines a multivariate criterion that is able to find all sparse loading vectors simultaneously. In computation, our algorithms developed in § 4 simplify greatly. In fact, because of the identity design, the inner loop in Step 5 of Algorithm 1 converges in one iteration and the overall procedure reduces to

$$XS^{(t-1)} = UDV^{\mathrm{T}}, \quad S^{(t)} \leftarrow \vec{\Theta}(X^{\mathrm{T}}UV^{\mathrm{T}}; \lambda). \tag{16}$$

In other words, $S^{(t)}$ is updated by thresholding $X^{\mathrm{T}}XS^{(t-1)}\{(S^{(t-1)})^{\mathrm{T}}X^{\mathrm{T}}XS^{(t-1)}\}^{-1/2}$, and various thresholding operators can be used. When $r = 1$, the singular value decomposition is unnecessary, since UV^{T} can be directly obtained by normalizing the column vector $XS^{(t-1)}$.

We now point out some recent literature related to (14) and (15). Under a spiked covariance model assumption, Cai et al. (2013) proposed and studied an adaptive multi-step procedure to solve a problem similar to (14). Our algorithm (16) is closest in spirit to the thresholding procedure in Johnstone & Lu (2009). Ma (2013) proposed an iterative algorithm for principal subspace estimation, but it has no guarantee of numerical convergence. Another type of sparse principal component analysis sets $Y = X$ in (11); the idea seems to have first appeared in Zou et al. (2006). But the self-regression formulation may bring some ambiguity in selection. Consider a noise-free model where $\mathcal{J}^*$ gives the set of indices of all nonzero columns in X, and $r^* = r(X)$ obeys $r^* < |\mathcal{J}^*|$. Then, given any index set $\mathcal{J} \subset \{1, \ldots, p\}$ satisfying $r(X_{\mathcal{J}}) = r^*$, we can find a matrix B with $\mathcal{J}(B) \subset \mathcal{J}$ and $r(B) \leqslant r^*$ such that $X = XB$, but $|\mathcal{J}|$ can be smaller or larger than $|\mathcal{J}^*|$.

6. APPLICATIONS

6·1. *Macroeconomic data*

Stock & Watson (2012) summarized 194 quarterly observations on 144 macroeconomic time series observed from 1960 to 2008, with some earlier observations used for lagged values of regressors as necessary. We pre-processed the data using the transformations given in Table B.2 of Stock & Watson (2012). One variable, non-borrowed reserves of depository institutions, was removed because its transformation involves logarithms but it has negative values. Of the 143 series, 35 are high-level aggregates, the information of which is all contained in the rest. Our predictors are the 108 disaggregated series and their lagged values. The series are grouped into 13 broad categories. We use the interest rates category, which consists of 13 time series, as our response variables. The dataset is a good example for showing that although forecasters and policymakers can access many potentially relevant macroeconometric time series, excluding noninformative ones is often ad hoc. In fact, to the best of our knowledge, there is no acknowledged model to describe all the 13 interest rates covering treasuries, corporate, term spreads and public-provide spreads. Low-rank models naturally arise (Reinsel & Velu, 1998), and the necessary factors can be as few as two or even one, which has been theoretically established in a large body of economics literature.

First, we used the 108 series observed in the past four quarters as predictors, giving $p = 4 \times 108 = 432$ predictors and 432×13 unknowns. We used Algorithm 1 for selective reduced rank regression and the scale-free predictive information criterion for parameter tuning. The resulting model has $\hat{r} = 1$ and $\hat{J} = 3$, achieving a remarkable dimension reduction. The single explanatory factor in response to all interest rate series is constructed from capital utilization, the three-month treasury bill secondary market rate minus the ten-year treasury constant maturity rate, and Moody's Baa corporate bond yield minus the ten-year treasury constant maturity rate. Since no variables of lag order 2 or above were selected, we repeated the analysis using the series with only one lag, thus selecting four variables and two factors. The last two variables shown in the one-factor model appear again, and the employment category contributes the other two variables, relating to employees on nonfarm payrolls in wholesale trade and help-wanted advertising in newspapers. Both the one-factor model and the two-factor model show a high level of parsimony.

Table 1. *Mean squared errors of autoregression with four lags, the one-factor model, and the two-factor model, for* 13 *time series in the interest rate category, with their medians and means reported in the last column*

Series index	1	2	3	4	5	6	7	8	9	10	11	12	13	(median, mean)
AR(4)	9·2	10·6	10·3	13·5	12·3	8·7	7·1	9·0	0·7	2·4	12·7	1·7	1·9	(9·0, 7·7)
1-factor	9·7	7·8	8·4	8·8	16·0	8·0	5·4	6·9	1·7	10·1	14·2	8·0	8·8	(8·4, 8·8)
2-factor	9·5	8·4	8·8	12·7	10·4	8·4	5·3	7·3	1·3	8·1	11·1	6·4	8·3	(8·3, 8·1)

AR(4), autoregression with four lags.

Next, we did a forecasting experiment to compare the obtained factor models with autoregressive modelling with four lags, which is a conventional but quite accurate forecasting method. The performance is evaluated by a rolling scheme: a rolling estimation window of the most recent 100 quarterly observations is used to estimate the parameters, and forecasts are made in the forecast window. Both windows move forward by one quarter at a time, and the procedure is repeated 94 times. Table 1 shows the mean squared errors of each method for the 13 interest rates. Overall, the three methods have comparable prediction errors. Of course, the comparison is, in some sense, unfair to factor models. The autoregression method builds a separate model for each interest rate using four relevant predictors, while the one-factor method, say, regresses every response on the same single score variable. Our purpose here is to demonstrate the usefulness of category-level factors; better models can possibly be built on the factors to improve the accuracy further.

6·2. *Face data*

The Extended Yale Face Database B (Lee et al., 2005; Georghiades et al., 2001) contains aligned and cropped face images of 38 subjects with the same frontal pose under 64 different illumination conditions. We down-sampled the images to 96×84, each containing 8064 pixels. Given a subject, a data matrix of size 64×8064 can be formed from the associated images. In face recognition, principal component analysis is widely used to extract basis features, referred to as eigenfaces, the number of which is controlled by the rank. We set $r = 30$ throughout this experiment and focus on the 22nd subject in the database, whose image examples are shown in the upper panels of Fig. 1. Selective principal component analysis was performed in the hope of capturing regions of interest under different light source directions. As seen in the lower panels, some informative regions sensitive to illumination conditions, e.g., forehead and nose tip, were automatically detected.

To reduce the computational burden caused by the large number of pixels, we tested the screening-guided hybrid sparse principal component analysis. See its description below (13) and recall that d^e controls the number of nonzero elements in the loading matrix S, and d controls the number of nonzero rows. Table 2 shows the results with a conservative screening choice $d = d^e$. The adjusted variance rates were computed according to Shen & Huang (2008). The hybrid principal component analysis gave essentially the same adjusted variances as sparse principal component analysis, but used fewer pixels. More importantly, the hybrid approach offered impressive time savings. Then, we did an experiment with a less conservative screening choice, $d = 2400$ and $d^e = 3600$. Sparse principal component analysis used 3517 pixels to reach an adjusted variance rate of 43%, while hybrid principal component analysis reduced the model size to 2400 pixels, and gave an adjusted variance rate of 40%. When using 2400 pixels, sparse principal component analysis only reached an adjusted variance rate of 34%.

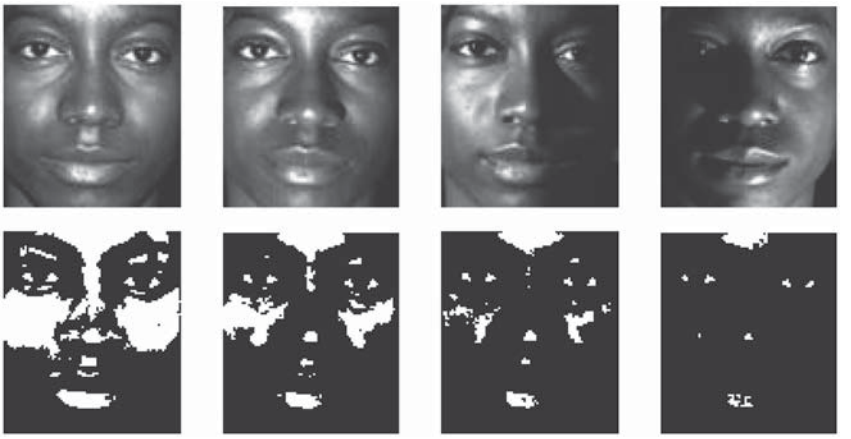

Fig. 1. The Upper row shows face image examples for subject 22 with image numbers $1, 10, 20$ and 50. The Lower row shows regions of interest, in white, identified by selective principal component analysis with $r = 30$ and $d = 2400, 1200, 600, 200$ from left to right.

Table 2. *Performance comparison between sparse and hybrid principal component analyses in terms of computational time, adjusted variance percentage, and number of pixels involved*

	$d^e = 4800$			$d^e = 2400$			$d^e = 1200$		
	Time	Pixels	AV	Time	Pixels	AV	Time	Pixels	AV
Sparse	254	4435	51·3	278	2370	33·4	293	1198	20·8
Hybrid	131	4377	51·1	87	2187	35·9	81	1127	20·5

AV, adjusted variance; Sparse, sparse principal component analysis; Hybrid, hybrid principal component analysis.

7. DISCUSSION

The techniques we have developed to study selective reduced rank regression are applicable to pure variable selection or rank reduction. For example, the recipe for proving Theorem 1 can handle Schatten p-norm penalized trace regression models, without using the sophisticated quasi-convex Schatten class embeddings (Rohde & Tsybakov, 2011). The scale-free predictive information criterion addresses the issue of adaptive rank selection in $p \gg n$ models, as raised by Bunea et al. (2011).

In this work, all the problems under consideration are nonconvex. In common with papers such as Rohde & Tsybakov (2011) and Bunea et al. (2012), we studied the properties of global minimizers. In some less challenging situations, the proposed algorithms, when initialized by the reduced rank regression estimator, can deliver a good estimate within a few iteration steps. This was also evidenced by Ma et al. in a recent technical report (arXiv:1403.1922). In some hard cases, we found the multi-start strategy of Rousseeuw & Van Driessen (1999) to be quite effective. The study of how to initialize and when to terminate is beyond the scope of the current paper, but is an interesting topic for further research.

ACKNOWLEDGEMENT

The author would like to thank the editor, associate editor and referees for their careful comments and useful suggestions that significantly improved the paper. The author also thanks

Florentina Bunea and Marten Wegkamp for helpful discussions. This work was supported in part by the National Science Foundation of the U.S.A.

Supplementary material

Supplementary material available at *Biometrika* online includes all technical details of the proofs and additional data analysis.

References

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

Berthet, A. & Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. *J. Mach. Learn. Res.* **30**, 1046–66.

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.

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

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

Cai, T. T., Ma, Z. & Wu, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41**, 3074–110.

Candès, E. J. & Plan, Y. (2011). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Info. Theory* **57**, 2342–59.

Chen, J. & Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika* **95**, 759–71.

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

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

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with Discussion). *J. R. Statist. Soc.* B **70**, 849–911.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–75.

Gao, C., Ma, Z. & Zhou, H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* to appear.

Georghiades, A., Belhumeur, P. & Kriegman, D. (2001). From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pat. Anal. Mach. Intel.* **23**, 643–60.

Hastie, T. J., Tibshirani, R. J. & Friedman, J. H. (2009). *The Elements of Statistical Learning*. New York: Springer, 2nd ed.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–41, 498–520.

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

Johnstone, I. M. & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc.* **104**, 682–93.

Lee, K., Ho, J. & Kriegman, D. (2005). Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pat. Anal. Mach. Intel.* **27**, 684–98.

Lounici, K., Pontil, M., Tsybakov, A. B. & van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39**, 2164–204.

Ma, X., Xiao, L. & Wong, W. H. (2014). Learning regulatory programs by threshold SVD regression. *Proc. Nat. Acad. Sci.* **111**, 15675–80.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41**, 772–801.

Recht, B., Fazel, M. & Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52**, 471–501.

Reinsel, G. & 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.

ROUSSEEUW, P. & VAN DRIESSEN, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–23.
SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comp. Statist. Data Anal.* **9**, 2976–90.
SHE, Y. (2016). On the finite-sample analysis of θ-estimators. *Electron. J. Statist.* **9**, 3098–119.
SHEN, H. & HUANG, J. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Mult. Anal.* **99**, 1015–34.
STOCK, J. H. & WATSON, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *J. Bus. Econ. Statist.* **30**, 481–93.
VAN DE GEER, S. A. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3**, 1360–92.
WEI, F. & HUANG, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16**, 1369–84.
WITTEN, D., TIBSHIRANI, R. J. & HASTIE, T. J. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *J. Mult. Anal.* **10**, 515–34.
YUAN, M. & LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.
ZHANG, C.-H. (2010a). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894–942.
ZHANG, C.-H. & ZHANG, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems. *Statist. Sci.* **27**, 576–93.
ZHANG, T. (2010b). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11**, 1081–107.
ZOU, H., HASTIE, T. J. & TIBSHIRANI, R. J. (2006). Sparse principal component analysis. *J. Comp. Graph. Statist.* **15**, 265–86.

[*Received on* 24 *March* 2014. *Editorial decision on* 24 *October* 2016]