Published in final edited form as:

J Stat Theory Pract. 2016 ; 10(1): 100–120. doi:10.1080/15598608.2015.1081573.

A note on rank reduction in sparse multivariate regression

Kun Chen^a and Kung-Sik Chan^b

^aDepartment of Statistics, University of Connecticut, Storrs, Connecticut, USA

^bDepartment of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa, USA

Abstract

A reduced-rank regression with sparse singular value decomposition (RSSVD) approach was proposed by Chen et al. for conducting variable selection in a reduced-rank model. To jointly model the multivariate response, the method efficiently constructs a prespecified number of latent variables as some sparse linear combinations of the predictors. Here, we generalize the method to also perform rank reduction, and enable its usage in reduced-rank vector autoregressive (VAR) modeling to perform automatic rank determination and order selection. We show that in the context of stationary time-series data, the generalized approach correctly identifies both the model rank and the sparse dependence structure between the multivariate response and the predictors, with probability one asymptotically. We demonstrate the efficacy of the proposed method by simulations and analyzing a macro-economical multivariate time series using a reduced-rank VAR model.

Keywords

Reduced-rank; rank selection; singular value decomposition; sparsity; vector autoregressive model

1. Introduction

In multivariate regression, a fundamental goal is to uncover meaningful dependence structure between the responses and the predictors. In many applications, (I) the responses and predictors may be associated through only a few latent channels/pathways, or (II) any relationship between the responses and predictors may only involve a subset of the responses and predictors. One approach of accounting for (I) is the reduced-rank regression (RRR) method (Anderson 1951; Izenman 1975; Reinsel and Velu 1998), which achieves dimension reduction through restricting the rank of the coefficient matrix. The rank indicates the number of latent predictors needed to model all the response variables, which are constructed as some linear combinations of the original predictors. Addressing (II), on the other hand, may require utilizing some sparsity-inducing regularization techniques such as lasso (Tibshirani 1996; Knight and Fu 2000) and adaptive lasso (Zou and Hastie 2005); see, for example, Witten et al. (2009), Lee et al. (2010), and Peng et al. (2010).

A more practical scenario is that both the dependence structures described in (I) and (II) are present simultaneously. Such complex dependence structures arise in many applications across disciplines, including ecology, cancer genetics, and genome-wide association studies. Chen et al. (2012) proposed a reduced-rank regression with sparse singular value decomposition (RSSVD) approach, which aims to identify the sparsity pattern in the singular value decomposition (SVD) of the coefficient matrix when its rank is correctly specified in advance. In RSSVD, each latent predictor is constructed from a subset of the original predictors due to predictor selection, and each response is allowed to be associated with possibly a subset of the constructed latent predictors. Assuming the true rank is known, the method was shown to enjoy the oracle properties (Knight and Fu 2000; Fan and Li 2001). Chen and Huang (2012) proposed a rank-constrained adaptive group lasso approach to recover a low-rank coefficient matrix with sparse rows for predictor selection (Yuan and Lin 2006; Huang et al. 2012a). A similar approach was also proposed by Bunea et al. (2012), who derived its nonasymptotic oracle bounds. Ma and Sun (2014) proposed a subspace assisted regression with row sparsity and studied its minimax optimality in estimation. These methods assume that the latent variables are constructed from the same subset of predictors, which is equivalent to imposing a special sparse SVD structure on the coefficient matrix, that is, the left singular vectors have the same zero entries and the right singular vectors need not be sparse. Therefore, the flexible RSSVD method can also be applied to recover the reduced-rank and row-sparse structure. Other recent work on the sparse SVD models include, for example, Chen et al. (2014), Zhu et al. (2014), and Ma et al. (2014).

It is clear that reduced-rank estimation is of intrinsic multivariate nature, and the three key ingredients in the regression analysis, that is, rank reduction, variable selection, and parameter estimation, are all interrelated. It is thus beneficial to conduct the three tasks simultaneously in such a way that they can be complementary to each other. Here, we show that a generalized RSSVD approach (Chen et al, 2012) is able to shrink any redundant singular component to a zero matrix, so that the true rank can be recovered. We show that asymptotically both the rank and the sparse SVD structure of the coefficient matrix can be correctly identified with probability approaching 1 as the sample size increases. The results are derived in the framework of stationary time series, which enable the usage of RSSVD in multivariate time-series modeling. Computationally, rank reduction is incorporated in the proposed iterative algorithm, thereby rendering the RSSVD more attractive for multivariate analysis. The empirical performance of the method is investigated under various simulation setups. We analyze a macro-economical time-series data set to showcase the effectiveness of the proposed method in reduced-rank vector autoregressive (VAR) modeling, for simultaneous variable selection, VAR order section, and rank determination.

2. Sparse SVD setup in multivariate regression

Given n observations of the response $\mathbf{y}_i \in \mathbb{R}^q$ and predictor $\mathbf{x}_i \in \mathbb{R}^p$, we consider the multivariate regression model,

$$\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E} \quad (1)$$

where $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$, $\mathbf{C}$ is a $p \times q$ coefficient matrix, and $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_n)^\top$ is a random $n \times q$ matrix; the error vectors are assumed to be independently and identically distributed (i.i.d.) with mean vector $E(\mathbf{e}_i) = \mathbf{0}$ and covariance matrix $\text{Cov}(\mathbf{e}_i) = \Sigma$, a $q \times q$ positive-definite matrix. We assume the variables are centered so that there is no intercept term.

The SVD of the coefficient matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^\top = \sum_{k=1}^r \mathbf{C}_k, \quad (2)$$

where $r = \min(p, q)$, $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_r)$ consists of orthonormal left singular vectors, $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_r)$ consists of orthonormal right singular vectors, $\mathbf{D} = \text{diag}(d_k)$ is an $r \times r$ diagonal matrix with nonnegative singular values $d_1 > \dots > d_{r^*} > 0$ and $d_{r^*+1} = \dots = d_r = 0$ on its diagonal, and $\mathbf{C}_k = d_k \mathbf{u}_k \mathbf{v}_k^\top$. The nonzero singular values are assumed to be distinct so that each of the r^* unit-rank layers $\mathbf{C}_k = d_k \mathbf{u}_k \mathbf{v}_k^\top$ is unique up to the signs of the singular vectors.

The preceding SVD representation could be used to parsimoniously reveal the complex dependence structures imposed by (I) and (II) in section 1. First, it shows that $\mathbf{C}$ is composed of r^* orthogonal unit-rank layers with decreasing singular values, that is, there are r^* distinct channels of decreasing importance relating the responses to the predictors.

Second, for each layer k , the elements in $\mathbf{u}_k$ can be viewed as the predictor effects because $\mathbf{X}\mathbf{u}_k$ creates a latent variable as a linear combination of the original predictors; the elements in $\mathbf{v}_k$ can be viewed as the response effects corresponding to the latent variable $\mathbf{X}\mathbf{u}_k$. In order to conduct variable selection and facilitate model interpretation, it is then desirable that the left and right singular vectors are sparse; that is, each relationship (pathway) relating the responses to the predictors may involve a few responses and predictors.

Assuming the rank of $\mathbf{C}$ has been correctly identified to be r^* , Chen et al. (2012) proposed the RSSVD approach, in which the coefficient matrix $\mathbf{C}$ was estimated by minimizing a penalized sum of squares criterion with respect to the triplets $(d_k, \mathbf{u}_k, \mathbf{v}_k)$ for $k = 1, \dots, r^*$. The penalty term for each triplet $(d_k, \mathbf{u}_k, \mathbf{v}_k)$ is proportional to the weighted l_1 norm of the outer product of $\mathbf{u}_k$ and $\mathbf{v}_k$, which induces sparsity in both of them. An iterative algorithm was proposed to conduct local estimation in the vicinity of some consistent estimator of $\mathbf{C}$. The requirement of strict orthogonality among the right (left) singular vectors is relaxed to accommodate the search of the sparsity pattern within each of the singular vectors, which improves searching efficiency, yet results in a regularized estimator that enjoys the oracle properties; see details in Chen et al. (2012). In the following, we consider several extensions and usages of the RSSVD approach to further expand its applicability.

3. Generalized RSSVD approach

3.1. The Generalized RSSVD Criterion

We generalize the RSSVD approach to conduct rank reduction, variable selection, and model estimation simultaneously. Although the true rank r^* is unknown, an upper bound of

the true rank is always available—for example, a trivial one is $r = \min(p, q)$; a sharper upper bound may be obtained in various ways, for example, based on some classical estimation methods or some substantive information of the underlying system. We propose to conduct sparse and reduced-rank estimation of $\mathbf{C}$ by minimizing

$$\frac{1}{2} \|\mathbf{Y} - \mathbf{X} \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^\top\|_F^2 + \sum_{k=1}^r \text{Pe}(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)), \quad (3)$$

with

$$\begin{aligned} \text{Pe}(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)) &= \lambda_k \left\| \left(w_k^{(d)} \mathbf{w}_k^{(u)} \mathbf{w}_k^{(v)\top} \right) \circ (d_k \mathbf{u}_k \mathbf{v}_k^\top) \right\|_1 \\ &= \lambda_k \sum_{i=1}^p \sum_{j=1}^q w_{ijk} |d_k u_{ik} v_{jk}| \\ &= \lambda_k (w_k^{(d)} d_k) \left(\sum_{i=1}^p w_{ik}^{(u)} |u_{ik}| \right) \left(\sum_{j=1}^q w_{jk}^{(v)} |v_{jk}| \right), \end{aligned} \quad (4)$$

where we require $\|\mathbf{u}_k\| = \|\mathbf{v}_k\| = 1$ with $\|\cdot\|$ denoting the l_2 norm; the notation $\|\cdot\|_F$ denotes the Frobenius norm, $\|\cdot\|_1$ the l_1 norm, and $\circ$ the Hadamard product or entrywise multiplication operator. The term $\sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^\top$ corresponds to the SVD of $\mathbf{C}$, and λ_k s are the tuning parameters. The weights $w_{ijk} = w_k^{(d)} w_{ik}^{(u)} w_{jk}^{(v)}$ are constructed from the rank- r reduced-rank estimator $\tilde{\mathbf{C}}^{(r)}$, obtained by minimizing the unpenalized sum of squares objective function (Reinsel and Velu 1998), that is,

$$\begin{aligned} w_k^{(d)} &= |\tilde{d}_k|^{-\gamma}, \\ w_{ik}^{(u)} &= |\tilde{u}_{ik}|^{-\gamma}, \\ \mathbf{w}_k^{(u)} &= \left(w_{1k}^{(u)}, \dots, w_{pk}^{(u)} \right)^\top, \\ w_{jk}^{(v)} &= |\tilde{v}_{jk}|^{-\gamma}, \\ \mathbf{w}_k^{(v)} &= \left(w_{1k}^{(v)}, \dots, w_{qk}^{(v)} \right)^\top, \end{aligned}$$

with $\tilde{\mathbf{C}}^{(r)} = \sum_k \tilde{d}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^\top = \sum_{k=1}^r \tilde{\mathbf{C}}_k^{(r)}$ being the SVD of $\tilde{\mathbf{C}}^{(r)}$ and γ being a prespecified constant that is set to be 2 in all the numerical works reported below (Zou 2006).

A main difference distinguishing the objective function (3) from the one proposed by Chen et al. (2012) is that Eq. (3) treats the model rank as unknown and thus it involves additional $r - r^*$ redundant SVD layers, of which the singular vectors are not identifiable. The redundancy is due to the uncertainty in the true rank, and this makes estimation and theoretical analysis challenging. The problem now involves both rank selection, that is,

determining the number of SVD layers or the number of nonzero singular values, and intrarelationship variable selection, that is, producing sparse singular vectors.

We show in section 3.2 that the redundant SVD layers can be successfully eliminated via the proposed regularization approach (3), and the proposed method enjoys both rank and variable selection consistency as well as asymptotic normality, that is, the oracle properties in the sense of Fan and Li (2001). Some insights on why the generalized approach works are in order. First, the penalty term $\text{Pe}(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k))$ defined in Eq. (4) is analogue to the adaptive lasso penalty (Zou 2006), and it is proportional to the product of a weighted singular value d_k , a weighted l_1 norm of $\mathbf{u}_k$, and a weighted l_1 norm of $\mathbf{v}_k$. Due to this multiplicative form, it penalizes the weighted l_1 norm of each SVD layer $d_k \mathbf{u}_k \mathbf{v}_k^\top$, which fairly distributes the penalties on the entries of each SVD layer and leads to automatic balance of the degree of sparsity between $\mathbf{u}_k$ and $\mathbf{v}_k$. Second, although the rank may be overspecified, we show that the initial rank- r reduced-rank estimator $\mathbf{C}^{(r)}$ and its singular values and singular vectors remain to be consistent for estimating their corresponding true counterparts (see Lemma 3.2.1), which justifies the validity of the adaptive penalization approach adopted in Eq. (3). Another important insight is that the penalty smoothly shrinks the singular value of each SVD layer toward 0. Although the singular vectors of the redundant layers are not identifiable, the penalty term for each $k > r^*$ promotes any redundant layer to become a zero matrix, due to its shrinkage and thresholding effects on the singular value. To see this, consider fixing $\mathbf{u}_k$ and $\mathbf{v}_k$ ($k = 1, \dots, r$). Then the penalty term in Eq. (3) can be viewed as a weighted l_1 norm of the singular values, that is, a weighted nuclear norm, which may induce sparsity among the singular values (Yuan et al. 2007; Chen et al. 2013). In fact, the objective function of Eq. (3) with fixed $\mathbf{u}_k$ and $\mathbf{v}_k$ ($k = 1, \dots, r$) can be written as an adaptive positive lasso problem (Efron et al. 2004) with respect to $\mathbf{d} = (\mathbf{d}_1, \dots, \mathbf{d}_r)^\top \geq \mathbf{0}$,

$$\frac{1}{2} \|\mathbf{y} - \mathbf{Z}\mathbf{d}\|^2 + \sum_{k=1}^r h_k d_k, \quad (5)$$

where $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_r)$ with $\mathbf{z}_k = \text{vec}(\mathbf{X}\mathbf{u}_k \mathbf{v}_k^\top)$, and $h_k = \lambda_k \sum_{i,j} w_{ijk} |u_{ik} v_{jk}|$. This suggests that the estimation of the singular values and hence the rank could be enhanced by incorporating Eq. (5) into optimization.

3.2. Theoretical properties

We follow the framework developed by Knight and Fu (2000), Fan and Li (2001), Zou (2006), and Chen et al. (2012) to study the theoretical properties of the generalized RSSVD estimator obtained by minimizing Eq. (3). We consider the general setting such that both the sparse SVD pattern and the model rank are unknown; the latter relaxation engenders several difficulties including the construction of consistent initial estimator and the identifiability issue of the singular vectors corresponding to the zero singular value. All proofs are given in the appendix.

We consider the following conditions.

C1. $\frac{1}{n} \mathbf{X}^\top \mathbf{X} \rightarrow \Gamma$ almost surely as $n \rightarrow \infty$, where Γ is a fixed, positive-definite matrix.

C2. The $\mathbf{e}_i (i = 1, \dots, n)$ are i.i.d. with $E(\mathbf{e}_i) = \mathbf{0}$ and $\text{Cov}(\mathbf{e}_i) = \Sigma$, a positive-definite matrix.

C3. $\frac{\lambda_k^{(n)}}{\sqrt{n}} \rightarrow 0$ and $\frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} \rightarrow \infty$ as $n \rightarrow \infty$, for $k = 1, \dots, r^*$, with $\gamma > 0$ being prespecified.

C4. $\frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} \rightarrow \infty$ as $n \rightarrow \infty$, for $k = r^* + 1, \dots, r$, with $\gamma > 0$ being prespecified.

We first define some notations. Let $\mathcal{A}$ and $\mathcal{B}$ be some subsets of the collection of row and column indices of an arbitrary matrix $\mathbf{Z}$, respectively. We let $\mathbf{Z}_{\mathcal{A}\mathcal{B}}$ denote a submatrix of $\mathbf{Z}$ whose rows and columns are chosen from $\mathbf{Z}$ according to the index sets $\mathcal{A}$ and $\mathcal{B}$, respectively. For simplicity, we may write $\mathbf{Z}_{\mathcal{A}\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a square matrix, $\mathbf{Z}_{\mathcal{A}\mathcal{B}} = \mathbf{Z}_{\mathcal{B}}(\mathbf{Z}_{\mathcal{A}})$ when $\mathcal{A}$ ($\mathcal{B}$) consists of all the row (column) indices, and $\mathbf{Z}_{\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}$ when $\mathbf{Z}$ is a vector.

Although the true rank r^* is unknown, some upper bound is always available, that is, $\text{rank}(\mathbf{C}) \leq r$ where $r^* \leq r \leq \min(p, q)$. The set of all $p \times q$ matrices of rank smaller than or

equal to r , denoted as $\Omega^{(r)}$, can be presented as a manifold that is a union of $\binom{q}{r}$ subsets each of which admits an Euclidean coordinate system, that is, $\Omega^{(r)} = \bigcup_{\mathcal{L} \in \Pi} \Omega_{\mathcal{L}}^{(r)}$, where Π consists of all size- r subsets of the set $\{1, \dots, q\}$ and

$$\Omega_{\mathcal{L}}^{(r)} = \{\mathbf{U}\mathbf{V}^\top; \mathbf{U} \in \mathbb{R}^{p \times r}, \mathbf{V} \in \mathbb{R}^{q \times r} \text{ with } \mathbf{V}_{\mathcal{L}} = \mathbf{I}_r\}.$$

Let $\mathbf{C} = \mathbf{U}^* \mathbf{V}^{*\top}$, where $\mathbf{U}_{p \times r}^* = (\mathbf{U}_{p \times r}^{**}, \mathbf{0}_{p \times (r-r^*)})$ with $\mathbf{U}_{p \times r^*}^{**} = (\mathbf{u}_1^*, \dots, \mathbf{u}_{r^*}^*)$ being a $p \times r^*$ orthogonal matrix, and $\mathbf{V}_{q \times r}^* = (\mathbf{V}_{q \times r^*}^{**}, \mathbf{V}_{q \times (r-r^*)}^{*\circ})$ is a $q \times r$ orthogonal matrix. Moreover, without loss of generality, we assume $\mathbf{C} \in \Omega_{\mathcal{L}}^{(r)} \subseteq \Omega^{(r)}$ where the index set $\mathcal{L} = \{l_1, \dots, l_r\}$ is arbitrary and fixed. This is equivalent to setting the submatrix $\mathbf{V}_{\mathcal{L}}^*$ of $\mathbf{V}^*$, also denoted as $\mathbf{Q}$, to be invertible. Denote $\check{\mathbf{U}} = \mathbf{U}^* \mathbf{Q}^\top$ and $\check{\mathbf{V}} = \mathbf{V}^* \mathbf{Q}^{-1}$, so that $\check{\mathbf{V}}_{\mathcal{L}} = \mathbf{I}_r$ and $\mathbf{C} = \mathbf{U}^* \mathbf{V}^{*\top} = \check{\mathbf{U}} \check{\mathbf{V}}^\top$. Henceforth, unless otherwise stated, the singular values are absorbed into the singular vectors for simplicity.

Let $Q_n(\mathbf{U}, \mathbf{V})$ denote the objective function, that is,

$$Q_n(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|_F^2 + \sum_{k=1}^r \left\{ \lambda_k^{(n)} \sum_{i=1}^p \sum_{j=1}^q w_{ijk} |u_{ik} v_{jk}| \right\}, \quad (6)$$

where $\mathbf{U} = (u_{ik})_{p \times r}$ and $\mathbf{V} = (v_{jk})_{q \times r}$. To find a local minimum of Eq. (6) in $\Omega_{\mathcal{L}}^{(r)}$, we consider a ball centered at $\mathbf{C}$ of radius $h > 0$:

$$\mathcal{N}(\mathbf{C}, h) = \left\{ \left(\hat{\mathbf{U}} + \frac{1}{\sqrt{n}} \check{\mathbf{A}} \right) \left(\hat{\mathbf{V}} + \frac{1}{\sqrt{n}} \check{\mathbf{B}} \right)^{\top}; \check{\mathbf{A}} = \begin{pmatrix} \check{a} \\ ik \end{pmatrix}_{p \times r} \text{ with } \|\check{\mathbf{A}}\| \leq h, \right. \\ \left. \check{\mathbf{B}} = (\check{b}_{jk})_{q \times r} \text{ with } \check{\mathbf{B}}_{\mathcal{L}} = \mathbf{0} \text{ and } \|\check{\mathbf{B}}\| \leq h, \right\} \cup \left\{ \left(\mathbf{U}^* + \frac{1}{\sqrt{n}} \check{\mathbf{A}} \right) \left(\mathbf{V}^* + \frac{1}{\sqrt{n}} \check{\mathbf{B}} \right)^{\top}; \check{\mathbf{A}} = (\check{a}_{ik})_{p \times r} \text{ with } \|\check{\mathbf{A}}\| \leq h, \right. \\ \left. \check{\mathbf{B}} = (\check{b}_{jk})_{q \times r} \text{ with } \check{\mathbf{B}}_{\mathcal{L}} = \mathbf{0} \text{ and } \|\check{\mathbf{B}}\| \leq h. \right\}.$$

where $\mathbf{A} = (a_{ik})_{p \times r} = \check{\mathbf{A}} \mathbf{Q}^{-\top}$ and $\mathbf{B} = (b_{jk})_{q \times r} = \check{\mathbf{B}} \mathbf{Q}$. A local minimum of $\mathcal{Q}_n(\mathbf{U}, \mathbf{V})$ is guaranteed to exist in $\mathcal{N}(\mathbf{C}, h)$ where the radius h is of the order $O(\sqrt{\log \log n})$, following similar arguments as in Chen et al. (2012, Theorem 1). We hence omit the details. Denote

the minimizer of Eq. (6) as $(\check{\mathbf{U}}, \check{\mathbf{V}})$ where $\hat{\mathbf{U}} = (\hat{u}_1, \dots, \hat{u}_r) = \left(\hat{\mathbf{U}}_{p \times r^*}^{(1)}, \hat{\mathbf{U}}_{p \times (r-r^*)}^{(2)} \right)$ and $\hat{\mathbf{V}} = (\hat{v}_1, \dots, \hat{v}_r) = \left(\hat{\mathbf{V}}_{q \times r^*}^{(1)}, \hat{\mathbf{V}}_{q \times (r-r^*)}^{(2)} \right)$.

We first show that when the model rank is overspecified, the initial rank- r reduced-rank estimator $\mathbf{C}^{(\tilde{r})}$, its singular values, and its first r^* singular vectors remain to be consistent for estimating their corresponding true counterparts. Proposition 3.2.1 gives the explicit form of the reduced-rank estimator $\mathbf{C}^{(\tilde{r})}$. We then show in Lemma 3.2.1 that the n -consistency of $\mathbf{C}^{(\tilde{r})}$ implies that $\tilde{d}_k$ is a n -consistent estimator of d_k , for $k = 1, \dots, r$, and so are $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{v}}_k$ for estimating $\mathbf{u}_k$ and $\mathbf{v}_k$, respectively, for $k = 1, \dots, r^*$. This is pivotal in computing the initial exclusive layers and the adaptive weights.

Proposition 3.2.1. (Reinsel and Velu 1998) The rank- r ($r \geq r^*$) estimator of $\mathbf{C}$ that minimizes $\|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_F^2$ subject to $\text{rank}(\mathbf{C}) = r$ is given by $\mathbf{C}^{(\tilde{r})} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{Y} \mathbf{H} \mathbf{H}^{\top}$, where $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_r]$ and $\mathbf{h}_k$ is the normalized eigenvector that corresponds to the k th largest eigenvalue of the matrix $\mathbf{Y}^{\top} \mathbf{X} (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{Y}$, for $k = 1, \dots, r$. Moreover, $\|\mathbf{C}^{(\tilde{r})} - \mathbf{C}\|_F = O_p(1)$.

Lemma 3.2.1. Let $\mathbf{C}^{(\tilde{r})}$ be the rank- r reduced-rank estimator of $\mathbf{C}$ as given in Proposition 3.2.1. Let $\mathbf{C} = \sum_{k=1}^r \mathbf{C}_k = \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^{\top}$ be the SVD of $\mathbf{C}$, where $d_1 > \dots > d_{r^*} > 0$ and $d_{r^*+1} = \dots = d_r = 0$. Similarly, let d_k , $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{v}}_k$, $k = 1, \dots, r$, be the singular values and left and right singular vectors, respectively, of $\mathbf{C}^{(\tilde{r})}$. Then under conditions C1 and C2,

- a. $n(\tilde{d}_k - d_k) = O_p(1)$, $n(\tilde{\mathbf{u}}_k - \mathbf{u}_k) = O_p(1)$ and $n(\tilde{\mathbf{v}}_k - \mathbf{v}_k) = O_p(1)$ for $k = 1, \dots, r^*$.
- b. $n\tilde{d}_k = O_p(1)$ for $k = r^* + 1, \dots, r$.

Our main results regarding the consistency of joint sparse SVD recovery and rank estimation are presented in the following two theorems.

Theorem 3.2.1. Suppose conditions C1-C4 are satisfied. Let $\hat{\mathbf{u}}_{\mathcal{A}}(u_{\mathcal{A}}^*)$ collect all entries in $\hat{\mathbf{U}}^{(1)}(\mathbf{U}^*)$ corresponding to the nonzero elements in $\mathbf{U}^*$, and let $\hat{\mathbf{v}}_{\mathcal{B}}(v_{\mathcal{B}}^*)$ collect all entries in $\hat{\mathbf{V}}^{(1)}(\mathbf{V}^*)$ corresponding to the nonzero elements in $\mathbf{V}^*$. Let $d_k = \|\hat{\mathbf{u}}_k\| \|\hat{\mathbf{v}}_k\|$ be the k th estimated singular values, for $k = r^* + 1, \dots, r$. Then

- a. $\sqrt{n}(\hat{\mathbf{u}}_{\mathcal{A}} - \mathbf{u}_{\mathcal{A}}^*)$ and $\sqrt{n}(\hat{\mathbf{v}}_{\mathcal{B}} - \mathbf{v}_{\mathcal{B}}^*)$ are both asymptotically normally distributed with zero mean; $\sqrt{n}(\hat{\mathbf{u}}_{\mathcal{A}^c} - \mathbf{u}_{\mathcal{A}^c}^*) \rightarrow_d 0$ and $\sqrt{n}(\hat{\mathbf{v}}_{\mathcal{B}^c} - \mathbf{v}_{\mathcal{B}^c}^*) \rightarrow_d 0$, as $n \rightarrow \infty$.
- b. $nd_k \rightarrow_d 0$, as $n \rightarrow \infty$, for $k = r^* + 1, \dots, r$.

Theorem 3.2.2. Suppose conditions C1-C4 are satisfied. Let $\mathcal{A}_k = \{i: u_{ik}^* \neq 0\}$, $\mathcal{B}_k = \{j: v_{jk}^* \neq 0\}$, $\mathcal{A}_k^{(n)} = \{i: \hat{u}_{ik} \neq 0\}$, and $\mathcal{B}_k^{(n)} = \{j: \hat{v}_{jk} \neq 0\}$, for $k = 1, \dots, r^*$. Let $d_k = \|\hat{\mathbf{u}}_k\| \|\hat{\mathbf{v}}_k\|$ be the k th estimated singular values, for $k = r^* + 1, \dots, r$. Then

- a. $P(\mathcal{A}_k^{(n)} = \mathcal{A}_k) \rightarrow 1$ and $P(\mathcal{B}_k^{(n)} = \mathcal{B}_k) \rightarrow 1$ as $n \rightarrow \infty$, for $k = 1, \dots, r^*$.
- b. $P(d_k = 0) \rightarrow 1$ as $n \rightarrow \infty$, for $k = r^* + 1, \dots, r$.

The preceding theoretical results assume i.i.d. data. In section 5, we consider a time-series application, in which the model is written as $\mathbf{y}_t = \mathbf{C}^\top \mathbf{x}_t + \mathbf{e}_t$ where $\mathbf{y}_t \in \mathbb{R}^q$, $\mathbf{x}_t \in \mathbb{R}^p$ and $\mathbf{e}_t \in \mathbb{R}^q$ is the error vector. The theoretical results can be extended to dependent data with additional conditions, that is:

C5. The process $\{(\mathbf{y}_t^\top, \mathbf{x}_t^\top)\}$ is stationary and ergodic. Let $\{\mathcal{F}_t\}$ be a sequence of increasing σ -algebras such that $\mathbf{x}_s, s \leq t+1$ and $\mathbf{e}_s, s \leq t$ are measurable w.r.t. $\mathcal{F}_t$. Moreover, $\{\mathbf{e}_t, \mathcal{F}_t\}$ is a martingale difference sequence, that is, $E(\mathbf{e}_t | \mathcal{F}_{t-1}) = 0$ a.s.

The preceding condition and the moment conditions stipulated in **C1** and **C2** then imply the validity of the law of large numbers, the law of iterated logarithm (Stout 2007), and the central limit theorem (Billingsley 1999, Theorem 18.3) that are needed for adapting the proofs of the theoretical results to dependent data. Theorems 3.2.1 and 3.2.2 show that the proposed method can identify the correct sparse and low-rank structure with probability approaching 1 as $n \rightarrow \infty$. We hence have demonstrated that the generalized RSSVD method achieves the oracle properties in both rank selection and variable selection.

3.3. Computation

Because of the nonconvexity of the objective function, we consider a local search strategy in the neighborhood of some consistent estimator of $\mathbf{C}$, for example, $\mathbf{C}^{(\tilde{r})}$ (see Lemma 3.2.1). We propose an extended iterative exclusive extraction algorithm (IEEA) to facilitate rank reduction. In the following, we use the notation IEEA(r) to indicate that the initial upper bound of the rank is r .

For $r = 1$, Eq. (3) reduces to the following problem:

$$\min_{(d, \mathbf{u}, \mathbf{v})} \left\{ \frac{1}{2} \|\mathbf{Y} - d\mathbf{X}\mathbf{u}\mathbf{v}^\top\|^2 + \lambda \sum_{i=1}^p \sum_{j=1}^q w_{ij} |du_i v_j| \right\}, \text{ s.t. } d \geq 0, \|\mathbf{u}\| = \|\mathbf{v}\| = 1, \quad (7)$$

where $w_{ij} = w_i^{(d)} w_i^{(d)} w_j^{(v)}$. Note that we have suppressed the subscript k for simplicity. The problem is “biconvex” in the sense that it becomes convex for either fixed $\mathbf{u}$ or $\mathbf{v}$. Therefore, the problem can be solved with a sparse unit-rank regression algorithm (SURR) (Chen et al. 2012). We present it here for the sake of completeness.

Sparse unit-rank regression algorithm (SURR)

1. Initialization step: Obtain an initial value $\mathbf{u}^{[0]}$ (or $\mathbf{v}^{[0]}$).
2. $\mathbf{v}$ -lasso step: Given $\mathbf{u}^{[i]}$, solve the following lasso problem w.r.t. $\check{\mathbf{v}} = \text{diag}(w^{(d)} \mathbf{w}^{(v)}) (d\mathbf{v})$:

$$\check{\mathbf{v}}^{[i]} = \arg \min_{\check{\mathbf{v}}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}^{(v)} \check{\mathbf{v}}\|^2 + \lambda^{(v)} \sum_{j=1}^q |\check{v}_j|,$$

where $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\mathbf{X}^{(v)} = \text{diag}(w^{(d)} \mathbf{w}^{(v)})^{-1} \otimes (\mathbf{X} \mathbf{u}^{[i]})$, and $\lambda^{(v)} = \lambda \left(\sum_{i=1}^p w_i^{(u)} |u_i^{[i]}| \right)$.

3. $\mathbf{v}$ -standardization step:

$$\tilde{d}^{[i+1]} = \|\text{diag}(w^{(d)} \mathbf{w}^{(v)})^{-1} \check{\mathbf{v}}^{[i]}\|, \quad \mathbf{v}^{[i+1]} = \text{diag}(w^{(d)} \mathbf{w}^{(v)})^{-1} \check{\mathbf{v}}^{[i]} / \tilde{d}^{[i+1]}.$$

4. $\mathbf{u}$ -lasso step: Given $\mathbf{v}^{[i+1]}$, solve the following lasso problem w.r.t. $\check{\mathbf{u}} = \text{diag}(w^{(d)} \mathbf{w}^{(u)}) (d\mathbf{u})$:

$$\check{\mathbf{u}}^{[i+1]} = \arg \min_{\check{\mathbf{u}}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}^{(u)} \check{\mathbf{u}}\|^2 + \lambda^{(u)} \sum_{i=1}^p |\check{u}_i|,$$

where $\mathbf{X}^{(u)} = \mathbf{v}^{[i+1]} \otimes (\mathbf{X} \text{diag}(w^{(d)} \mathbf{w}^{(u)})^{-1})$ and $\lambda^{(u)} = \lambda \left(\sum_{j=1}^q w_j^{(v)} |v_j^{[i+1]}| \right)$.

5. $\mathbf{u}$ -standardization step:

$$d^{[i+1]} = \|\text{diag}(w^{(d)} \mathbf{w}^{(u)})^{-1} \check{\mathbf{u}}^{[i+1]}\|, \quad \mathbf{u}^{[i+1]} = \text{diag}(w^{(d)} \mathbf{w}^{(u)})^{-1} \check{\mathbf{u}}^{[i+1]} / d^{[i+1]}.$$

6. For $i = 1, 2, \dots$, iterate through steps 2 to 5 until some convergence criterion is met, for example, $\|d^{[i+1]} \mathbf{u}^{[i+1]} \mathbf{v}^{[i+1]^\top} - d^{[i]} \mathbf{u}^{[i]} \mathbf{v}^{[i]^\top}\|_F < \epsilon$, where ϵ is a small positive number, for example, $\epsilon = 10^{-4}$.

The SURR algorithm is a blockwise coordinate descent method with two blocks. The lasso steps can be solved efficiently by various methods (Efron et al. 2004; Friedman et al. 2007). Due to the multiplicative form of the penalty term, the in-between lasso steps 2 and 4 are with respect to either weighted $d\mathbf{v}$ or weighted $d\mathbf{u}$ and hence the l_2 norm constraints on either $\mathbf{v}$ and $\mathbf{u}$ are avoided; more importantly, the in-between standardization steps 3 and 5 make the estimates of $\mathbf{u}$ and $\mathbf{v}$ satisfy the norm constraints without changing the value of the objective function. Therefore, the objective value decreases monotonically along the iterations, and the algorithm is hence stable and always converges. See Chen et al. (2012) for more details.

In our numerical work, we compute solutions along a grid of 100 λ values that are equally spaced on the log scale. Let $(\hat{d}^{(\lambda)}, \hat{\mathbf{u}}^{(\lambda)}, \hat{\mathbf{v}}^{(\lambda)})$ be the fitted value of $(d, \mathbf{u}, \mathbf{v})$ with the tuning parameter being λ . To select the best solution, we consider a Bayesian information criterion (Schwarz 1978; An et al. 2008),

$$\text{BIC}(\lambda) = \begin{cases} \log(\text{SSE}(\lambda)) + \frac{\log(qn)}{qn} \text{df}(\lambda) & \text{when } p < n \\ \log(\text{SSE}(\lambda)) + \frac{2\log(pq)}{qn} \text{df}(\lambda) & \text{when } p \geq n \end{cases} \quad (8)$$

where $\text{SSE}(\lambda) = \|\mathbf{Y} - \hat{d}^{(\lambda)} \mathbf{X} \hat{\mathbf{u}}^{(\lambda)} \hat{\mathbf{v}}^{(\lambda)\top}\|_F^2$ stands for the sum of squared error, and $\text{df}(\lambda) = \sum_{i=1}^p \mathbf{I}(\hat{u}_i^{(\lambda)} \neq 0) + \sum_{j=1}^q \mathbf{I}(\hat{v}_j^{(\lambda)} \neq 0) - 1$ is the estimated degrees of freedom of the model (Zou et al. 2007).

For the general case $r > 1$, the problem can be solved by iteratively conducting SURR problems. Since the rank is unknown, a rank or singular value estimation step can be incorporated. Starting with an initial estimator $\tilde{\mathbf{C}}^{(r)}$, define the exclusive layers

$\mathbf{Y}_k = \mathbf{Y} - \mathbf{X}(\tilde{\mathbf{C}}^{(r)} - \tilde{\mathbf{C}}_k^{(r)})$ for $k = 1, \dots, r$. Then fixing all the triplets $(d_i, \mathbf{u}_i, \mathbf{v}_i)$ but the k th one, the optimization of $(d_k, \mathbf{u}_k, \mathbf{v}_k)$ becomes a SURR problem of $\mathbf{Y}_k$ on $\mathbf{X}$, that is,

$\frac{1}{2} \|\mathbf{Y}_k - d_k \mathbf{X} \mathbf{u}_k \mathbf{v}_k^\top\|_F^2 + \text{Pe}(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k | k))$. Solving the SURR problems may yield sparse or even zero singular vector estimates, in which case the specified rank can be reduced.

Theoretically, it can be verified that such one-step estimators also achieve the oracle properties as established in the previous section (Chen 2011). In practice, to enhance the rank estimation, an extra step of updating the singular values and the rank based on solving Eq. (5) is then followed. We then iterate these SURR and rank refinement steps a number of times or until convergence, each time using the previous estimates as initial values.

Following Lee et al. (2010), the selection of the regularization parameters may be nested within the iterative algorithm, thereby avoiding the use of computationally expensive grid search of the r regularization parameters. We hence propose the following extended IEEA algorithm to minimize Eq. (3).

(Extended) Iterative exclusive-extraction algorithm (IEEA)

1. Initialization step: start from some initial estimator $\mathbf{C}^{[0]}$ with $r^{[0]} = r$, and its (estimated) SVD $\mathbf{C}^{[0]} = \sum_{k=1}^{r^{[0]}} d_k^{[0]} \mathbf{u}_k^{[0]} \mathbf{v}_k^{[0]\top} = \sum_{k=1}^{r^{[0]}} \mathbf{C}_k^{[0]}$.
2. EEA step:
 - a. Form the exclusive layers $\mathbf{Y}_k^{[i]} = \mathbf{Y} - \mathbf{X} (\mathbf{C}^{[i]} - \mathbf{C}_k^{[i]})$ for $k = 1, \dots, r^{[i]}$.
 - b. For each $k \in \{1, \dots, r^{[i]}\}$, find the solution path of the SURR problem of $\mathbf{Y}_k^{[i]}$ on $\mathbf{X}$; the optimal $\lambda_k^{[i+1]}$ and hence the solution $(\hat{d}_k^{[i+1]}, \mathbf{u}_k^{[i+1]}, \mathbf{v}_k^{[i+1]})$ can be chosen by either cross validation or the BIC defined in Eq. (8). Let $\tilde{r}^{[i+1]} = \sum_{k=1}^{r^{[i]}} I(\hat{d}_k^{[i+1]} \neq 0)$ and eliminate the estimated zero SVD layers.
3. Rank refinement step: For fixed $\mathbf{u}_k^{[i+1]}, \mathbf{v}_k^{[i+1]}$, and $\lambda_k^{[i+1]}$ for $k = 1, \dots, \tilde{r}^{[i+1]}$, solve Eq. (5) to obtain refined singular values $\mathbf{d}^{[i+1]}$. Let $r^{[i+1]} = \sum_{k=1}^{\tilde{r}^{[i+1]}} I(d_k^{[i+1]} \neq 0)$ and eliminate the estimated zero SVD layers.
4. For $i = 1, 2, \dots$, repeat steps 2 and 3 until some convergence criterion is met, for example, $r^{[i+1]} = r^{[i]}$ and $\|\mathbf{C}^{[i+1]} - \mathbf{C}^{[i]}\|_F / \|\mathbf{C}^{[i]}\|_F < \epsilon$ ($\epsilon = 10^{-4}$).

The preceding algorithm is efficient and stable, based on our limited experience. We have implemented the methods in R (R Development Core Team 2014), and C functions were linked and called at run time for the computationally-intensive parts. (An R package is available upon request.)

4. Simulation

We consider four simulation models with various model dimensions and sparse structures. In each model, the data matrix $\mathbf{Y}$ is generated as $\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}$, where the elements of $\mathbf{E}$ are i.i.d. samples from $N(0, \sigma^2)$, and σ is chosen to control the signal-to-noise ratio (SNR)

computed as $\|d_{r^*} \mathbf{X} \mathbf{u}_{r^*} \mathbf{v}_{r^*}^\top\|_F^2 / \|\mathbf{E}\|_F^2$.

Models I and II are from section 4.2 of Chen et al. (2012). The covariate matrix $\mathbf{X}$ is constructed by generating its n rows as i.i.d. samples from $N(0, \Gamma)$, where $\Gamma = (\Gamma_{ij})_{p \times p}$ and $\Gamma_{ij} = 0.5^{|i-j|}$. In Model I, we set $p = q = 25$, $n = 50$, and $r^* = 3$. Let $\mathbf{C}$ be a 25×25 rank-3

matrix whose SVD is $\sum_{k=1}^3 d_k \mathbf{u}_k \mathbf{v}_k^\top$ with $d_1 = 20$, $d_2 = 10$, and $d_3 = 5$. The $\mathbf{u}_k$ s and $\mathbf{v}_k$ s are given by

$$\begin{aligned}
 \check{\mathbf{u}}_1 &= [\text{unif}(\mathcal{A}_u, 5), \text{rep}(0, 20)]^\top, & \check{\mathbf{u}}_2 &= [\text{rep}(0, 5), \text{unif}(\mathcal{A}_u, 5), \text{rep}(0, 15)]^\top, \\
 \check{\mathbf{u}}_3 &= [\text{rep}(0, 10), \text{unif}(\mathcal{A}_u, 5), \text{rep}(0, 10)]^\top, & \mathbf{u}_k &= \check{\mathbf{u}}_k / \|\check{\mathbf{u}}_k\| \text{ for } k=1, 2, 3; \\
 \check{\mathbf{v}}_1 &= [\text{unif}(\mathcal{A}_v, 10), \text{rep}(0, 15)]^\top, & \check{\mathbf{v}}_2 &= [\text{rep}(0, 12), \text{unif}(\mathcal{A}_v, 10), \text{rep}(0, 3)]^\top, \\
 \check{\mathbf{v}}_3 &= [\text{rep}(0, 6), \check{\mathbf{v}}_{1,7:8}, -\check{\mathbf{v}}_{1,9:10}, \text{unif}(\mathcal{A}_v, 2), -\check{\mathbf{v}}_{2,13:14}, \check{\mathbf{v}}_{2,15:16}, \text{rep}(0, 9)]^\top, \\
 & & \mathbf{v}_k &= \check{\mathbf{v}}_k / \|\check{\mathbf{v}}_k\| \text{ for } k=1, 2, 3,
 \end{aligned}$$

where the notation $\text{unif}(\mathcal{A}, b)$ denotes a vector of length b whose entries are i.i.d. samples from the uniform distribution on the set of real values $\mathcal{A}$, $\mathcal{A}_u = [-1, -0.3] \cup [0.3, 1]$, $\mathcal{A}_v = \pm 1$, the notation $\text{rep}(a, b)$ denotes a vector of length b , whose entries are all equal to a , and $\check{\mathbf{v}}_{k,a:b}$ denotes the a th to b th entries of $\check{\mathbf{v}}_k$. In Model II, we set $p = q = 60$, $n = 50$, and $r^* = 3$. The setup is the same as in Model I, except that each 25×1 singular vector in Model I is appended with 35 zeros to make a 60×1 vector.

Models III and IV are similar to the setups in section 6.3 of Chen and Huang (2012). The rows of $\mathbf{X}$ are generated from $N(0, \Gamma)$, where Γ has diagonal elements 1 and off-diagonal elements 0.5. The coefficient matrix $\mathbf{C}$ is generated as $\mathbf{C} = \mathbf{B}\mathbf{A}^\top$. For $\mathbf{B} \in \mathbb{R}^{p \times r^*}$, the entries of its first p_0 rows are generated from $N(0, 1)$ and the rest of the entries are set to be zero; For $\mathbf{A} \in \mathbb{R}^{q \times r^*}$, its entries are all generated from $N(0, 1)$. In Model III, we set $p = 25$, $p_0 = 10$, $q = 10$, $n = 50$, and $r^* = 3$. In Model IV, the only change is $p = 60$; that is, the 25×3 matrix $\mathbf{B}$ in Model III is appended with 35 zero rows to form a 60×3 matrix.

In Models I and II, the left and right singular matrices of $\mathbf{C}$ are entrywise sparse. Thus, the models have exactly the sparse and low rank structure that the RSSVD method aims to recover. In Models III and IV, $\mathbf{C} = \mathbf{B}\mathbf{A}^\top$ is a low rank and row-sparse matrix, which means its left singular matrix is row sparse and its right singular matrix is not sparse. This is because the row sparsity of $\mathbf{B}$ is invariant to nonsingular transformation; that is, for any $r^* \times r^*$ invertible matrix $\mathbf{Q}$, let $\tilde{\mathbf{B}} = \mathbf{B}\mathbf{Q}$ and $\tilde{\mathbf{A}}^\top = \mathbf{Q}^{-1}\mathbf{A}^\top$, then $\mathbf{C} = \tilde{\mathbf{B}}\tilde{\mathbf{A}}^\top$ and $\tilde{\mathbf{B}}$ is still row sparse. This setup assumes that all the response variables are associated with the same subset of predictors.

We mainly consider the ordinary least squares estimator (OLS), the reduced-rank regression estimator (RRR), and the generalized RSSVD estimator. For the RSSVD method, we perform model estimation with initial rank $r = 8$ in Models I and II and $r = 5$ in Models III and IV. In practice a good upper bound of the true rank can be readily obtained by either performing a classical reduced-rank regression analysis or using a scree plot as in the principal component analysis. The tuning parameters are chosen by Bayesian information criterion (BIC). For Models III and IV, we also construct an easily computable oracle estimator, which is obtained by performing the RRR method using only the important p_0 predictors. The experiment was replicated 300 times for each SNR level. We report the percentage of correct rank determination (Rank); the model accuracy is measured by the average scaled mean squared error (SMSE) from all runs, that is, $\text{SMSE} = 100\text{tr}\{(\mathbf{C} - \hat{\mathbf{C}})^\top \Gamma (\mathbf{C} - \hat{\mathbf{C}})\}/(nq)$, where $\text{tr}(\cdot)$ denotes the trace of a matrix. The overall false discovery rates (FDR) and false negative rates (FNR) of the RSSVD method are calculated based on the estimated left and right singular matrices and their corresponding true counterparts. (The redundant singular vectors are considered as zero vectors.)

Tables 1 and 2 summarize the simulation results for Models I–II and III–IV, respectively. The generalized RSSVD method can indeed conduct automatic rank reduction and sparse SVD estimation, and the percentages of correct rank determination are generally high. The RRR method shows near-perfect rank determination performance while the performance of RSSVD method on rank determination gradually improves as noise level reduces. The main difference of the two methods on rank reduction is that the RRR method corresponds to an

l_0 -type singular value penalization, while the RSSVD method corresponds to an adaptive l_1 -type singular value penalization. Our results are consistent with Chen et al. (2013), in which it was shown that in the context of reduced-rank estimation, while both the l_0 and the adaptive l_1 singular value penalization methods can achieve rank selection consistency and the minimax rate for rank reduction; in practice, the adaptive l_1 singular value penalized estimator tends to slightly overestimate the rank. Examining the estimated singular values of RSSVD reveals that the redundant SVD layers, if not eliminated entirely, are usually shrunk to be nearly a zero matrix, so that overfitting is kept minimal. In all the cases, the RSSVD method greatly outperforms the RRR method and OLS method in terms of having much lower SMSE. This is mainly due to its capability of conducting simultaneous variable selection and rank reduction.

We also find that in Models I and II, the performance of the generalized RSSVD is, as expected, slightly worse than the case of known rank. Also, in Models III and IV, a better performance can be achieved if we assume the true row-sparse and reduced-rank structure is known. Nevertheless, without assuming the row-sparse structure, the RSSVD is still able to dramatically outperform RRR and its performance can be close to that of the oracle estimator (the RRR estimator assuming known row-sparsity pattern) when the signal is moderate to high. The generalized RSSVD relies on a smooth l_1 -type penalty to recover the rank; which is more stable than the l_0 -type rank constrained approach, and the method adapts to flexibly recover specific sparse structures such as row-sparsity.

We have also investigated the impact of initial rank specification on the performance of RSSVD and its computation speed, as suggested by a referee. The structure of the simulation model is the same as that of Model II where $(p, q, n) = (60, 60, 50)$, except that we consider different numbers of predictors $p \in \{75, 150, 300\}$. That is, each left singular vector in Model II is appended with $p - 60$ zeros to form a new $p \times 1$ singular vector. Recall that the true rank is $r^* = 3$. For the initial rank specification, we consider $r \in \{1, 2, \dots, 10\}$. To be consistent, in every SURR subroutine, the model is fitted with 100 tuning parameters and the BIC is used for tuning parameter selection. The experiment was replicated 100 times under each setting. All the computation was done on a computer with an Intel 2.9-GHz central processing unit (CPU) and 8GB RAM. In Figure 1a, we report the average computation times of RSSVD (total time using 100 tuning parameters). For each fixed p , the computation time increases linearly with the initial rank r . This is expected, as the RSSVD method involves estimating the r SVD layers using sparse unit-rank regression. In sparse high-dimensional problems, the true rank is usually small, and severe overestimation of the true rank is unlikely. Nevertheless, RSSVD is computationally efficient: Even when $p = 300$ and $r = 10$, the RSSVD analysis only took about 8 seconds on average. In Figure 1b, we report the average SMSEs of RSSVD; the SMSEs of the classical RRR when $p = 75$ are included as benchmarks. (For better illustration, the SMSEs of RRR for $p = 150/p = 300$ are omitted as these values are even higher than those for $p = 75$.) In all the cases, RSSVD offered great improvement over RRR. As expected, RSSVD with the true rank value yielded the lowest SMSE on average. However, as long as the rank was not underspecified, RSSVD with an initially misspecified rank value performed almost as well as the case of using the true rank. This is due to the capability of RSSVD for conducting automatic rank reduction.

5. Reduced-rank VAR modeling via RSSVD

Given n observations of a multivariate time-series response $\mathbf{y}_t \in \mathbb{R}^q$ ($t = 1, \dots, n$), consider a vector autoregressive (VAR) process of order h :

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \dots + \mathbf{A}_h \mathbf{y}_{t-h} + \mathbf{e}_t = \mathbf{A} \mathbf{x}_{t-1} + \mathbf{e}_t, \quad (9)$$

where $\mathbf{A} = (\mathbf{A}_1, \dots, \mathbf{A}_h)$ is an $hq \times q$ matrix consisting of all the $q \times q$ matrices $\mathbf{A}_i$ ($i = 1, \dots, h$), $\mathbf{x}_t = (\mathbf{y}_t^\top, \dots, \mathbf{y}_{t-h+1}^\top)^\top$, and the error $\mathbf{e}_t$ ($t = 1, \dots, n$) is a zero-mean innovation process with which condition **C5** in Remark 1 is assumed to hold. We assume the process is mean-adjusted so that there is no intercept term. Let $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top$, $\mathbf{X} = (\mathbf{x}_0, \dots, \mathbf{x}_{n-1})^\top$, $\mathbf{C} = \mathbf{A}^\top$, and $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_n)^\top$. Then model (9) can be written in standard multivariate regression form as $\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{E}$. (We assume h presample values are available.)

Upon assuming $\mathbf{C}$ to be a reduced-rank matrix of rank r^* , model (9) is then called a reduced-rank VAR model (Lütkepohl 1993). The coefficient matrix $\mathbf{C}$ is of reduced rank, which means that it can be written as $\mathbf{C} = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{U}$ is $hq \times r^*$ and $\mathbf{V}$ is $q \times r^*$. The rationale behind this formulation is that, except for multiplicative constants contained in $\mathbf{V}$, there may be only a few linear combinations of the $\mathbf{x}_{t-1}$ vector that contain all the relevant information regarding $\mathbf{y}_t$. This model is widely used in time-series analysis. A reduced-rank VAR model analysis generally consists of three parts, that is, the determination of the VAR order h , the determination of the rank r^* , and the estimation of the coefficient matrix $\mathbf{C}$.

The order determination problem is related to the more general problem of identifying the nonzero components in a VAR model. For seasonal time series, a VAR model of a high order h may be needed to capture the seasonality. Often, only lags in a subset between 1 and h are statistically significant. Finding a subset VAR model that adequately approximates the underlying process is more important from a practical standpoint than simply determining the VAR order (Chen and Chan 2011). The multiple fitting procedure, which relies on exhaustive subset model selection, may be computationally intensive and even impractical when h is large. Moreover, the variable association characterized by each reduced-rank layer of a VAR model may involve a subset of the variables and their lags but not all of them. Therefore, imposing the sparsity conditions on both $\mathbf{U}$ and $\mathbf{V}$ may enhance model interpretability and results in improved predictive accuracy through flexible variable and lag selection. It is then desirable to use the RSSVD approach to conduct the reduced-rank VAR model analysis.

We consider a three-dimensional time-series dataset $\{\mathbf{y}_t = (y_{1,t}, y_{2,t}, y_{3,t})^\top\}$ consisting of first differences of the logarithms of quarterly, seasonally adjusted West German fixed investment $\{y_{1,t}\}$, disposable income $\{y_{2,t}\}$, and consumption expenditures $\{y_{3,t}\}$ from 1960 to 1982. The trend in the original data is removed by taking first differences of logarithmic transformation, and the means of the variables are then subtracted. The original data and first differences of logarithms are plotted in Figure 2. Lütkepohl (1993) analyzed the data from 1960 to 1978 using reduced-rank VAR models, in which the rank- r estimator of $\mathbf{C}$ is obtained by minimizing $\text{tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{C})\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{C})^\top\}$ subject to $\text{rank}(\mathbf{C}) = r$, where

$\tilde{\Sigma} = \frac{1}{n} \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) \mathbf{Y}$. To be consistent with their analysis, we then use the transformed data $\tilde{\mathbf{Y}} = \mathbf{Y} \tilde{\Sigma}^{-1/2}$ in the RSSVD estimation.

Lütkepohl (1993) considered models of ranks $r = 1, 2, 3$ and VAR orders $h = 0, 1, 2, 3, 4$ (13 models in total). Their best model was selected based on the Akaike information criterion (AIC), which turned out to be a rank-1 VAR(2) model. We thus conduct the RSSVD estimation with $h = 4$ and $r = 3$. The regularization parameters are chosen by minimizing the BIC defined in Eq. (8). The RSSVD method is able to reduce the rank automatically, and the resulting estimate, denoted as $\hat{\mathbf{C}}_{SS}$, is of rank-1. Note that $\mathbf{C} = (\mathbf{A}_1, \dots, \mathbf{A}_4)^\top$ when $h = 4$. Interestingly, both $\mathbf{A}_3$ and $\mathbf{A}_4$ are estimated to be exactly zero matrices while $\mathbf{A}_1$ and $\mathbf{A}_2$ are not, which suggests that the underlying process may be more appropriately modeled by an VAR(2) model. Therefore, we refit the data by the RSSVD method with $h = 2$ and $r = 3$. The method again yields a sparse coefficient matrix of rank-1, and the fitted VAR(2) model can be written as

$$\mathbf{y}_t = \begin{pmatrix} 0 & 0.06 & -0.09 \\ 0 & -0.08 & 0.12 \\ 0 & 0.26 & -0.40 \end{pmatrix} \mathbf{y}_{t-1} + \begin{pmatrix} 0 & 0.06 & 0 \\ 0 & -0.08 & 0 \\ 0 & 0.28 & 0 \end{pmatrix} \mathbf{y}_{t-2} + \hat{\mathbf{e}}_t. \quad (10)$$

Our estimation result is consistent with that reported in Lütkepohl (1993); that is, both methods choose a rank-1 VAR(2) model, and the nonzero entries in $\hat{\mathbf{C}}_{SS}$ agree in sign with the corresponding entries in the estimate obtained in Lütkepohl (1993), denoted as $\hat{\mathbf{C}}_{HL}$. Several additional insights of the underlying system can be gained by examining the sparsity pattern in $\hat{\mathbf{C}}_{SS}$. It suggests that $\mathbf{y}_t$ may not be related to $y_{1,t-1}$, $y_{1,t-2}$, and $y_{3,t-2}$, given other covariates in the VAR(2) model—see Eq. (10), which, moreover, reveals that $\{\mathbf{w}_t = (y_{2,t}, y_{3,t})^\top\}$, that is, disposable income and consumer expenditure, forms a closed system following a VAR(2) model, and that the lags 1 and 2 of $\mathbf{w}_t$ drive $y_{1,t}$ (fixed investment).

We have also calculated the scaled within-sample prediction errors defined as

$1000 \|\mathbf{Y} - \mathbf{X} \hat{\mathbf{C}}\|_F^2 / (nq)$ using data from 1960 to 1978, and then the scaled out-of-sample prediction errors based on data from 1979 to 1982, for the LS estimator ($\hat{\mathbf{C}}_{LS}$), the estimator in Lütkepohl (1993) ($\hat{\mathbf{C}}_{HL}$), and the RSSVD estimator ($\hat{\mathbf{C}}_{SS}$). The within-sample errors are 0.710, 0.790 and 0.810 and the out-of-sample errors are 0.429, 0.308, and 0.299, respectively, for $\hat{\mathbf{C}}_{LS}$, $\hat{\mathbf{C}}_{HL}$, and $\hat{\mathbf{C}}_{SS}$. While the within-sample error of the RSSVD estimator is slightly higher than those of the other estimators, it performs slightly better in terms of having the lowest out-of-sample error.

6. Discussion

We have generalized the RSSVD approach to provide a unified multivariate regression framework for simultaneous rank reduction, variable selection, and model estimation. The reduced-rank regression estimator was used as the initial estimator for conducting adaptive regularized estimation, and the method is shown to be robust against overspecification of the initial rank. In practice, there are many ways of obtaining a good initial reduced-rank

estimator. In particular, the reduced-rank estimation can be achieved by penalizing the singular values of the coefficient matrix, and the rank selection consistency of these methods has been extensively studied; see, for example, Bunea et al. (2011) and Chen et al. (2013), among others. These reduced-rank estimators usually admit explicit solutions and are easy to compute.

Some pressing problems for future research include investigating the feasibility of sparse and (exactly) orthogonal SVD estimation in regression problems and extending the theory to high-dimensional settings with a diverging number of predictors and responses. We have suggested using $\gamma = 2$ to construct the adaptive weights, following Zou and Hastie (2005). In practice, the choice of γ can be data driven, which may be particularly beneficial when the quality of the initial estimator is uncertain. In many applications, the data may consist of more than one set of regressors, which may be addressed by considering RSSVD models with multiple sets of regressors (Reinsel and Velu 1998). Further, the reduced-rank structure may be distorted by the presence of outliers, and hence there is a strong need for developing a robust reduced-rank method (Mukherjee and Zhu 2011). Finally, it is pressing to extend the methodology to vector generalized linear models (Yee and Hastie 2003; She 2013), as well as semiparametric and nonparametric settings (Li and Chan 2007; Yuan et al. 2007; Huang et al. 2012b).

Acknowledgments

Funding: We gratefully acknowledge partial support from the National Science Foundation (DMS-1021896), the National Institutes of Health (NIH U01 HL114494), and the Simons Foundation (359404).

Appendix: Proofs

Proof of Lemma 3.2.1. With the consistency result in Lemma 3.1, the claims stated in (a) can be proved by techniques based on perturbation expansion of matrices; see, for example, Izenman (1975, Theorem 3), Bura and Pfeiffer (2008, Theorem 1), and Chen et al. (2012,

Lemma 1.2). Denote $\tilde{\Sigma}_{yx} = \tilde{\Sigma}_{xy}^\top = \frac{1}{n} \mathbf{Y}^\top \mathbf{X}$, $\tilde{\Sigma}_{xx} = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$, $\tilde{\mathbf{C}} = \tilde{\Sigma}_{xx}^{-1} \tilde{\Sigma}_{xy}$, and $\tilde{\Sigma}_{\varepsilon\varepsilon} = \frac{1}{n} (\mathbf{Y} - \mathbf{X}\tilde{\mathbf{C}})^\top (\mathbf{Y} - \mathbf{X}\tilde{\mathbf{C}})$. Let $\check{d}_k^2$ be the k th largest eigenvalue of the matrix $\tilde{\Sigma}_{\varepsilon\varepsilon}^{-1/2} \tilde{\Sigma}_{yx} \tilde{\Sigma}_{xx}^{-1} \tilde{\Sigma}_{xy} \tilde{\Sigma}_{\varepsilon\varepsilon}^{-1/2}$ for $k = 1, \dots, r$. Following Hsu (1941) and Anderson (1951, Theorem 3), we have

$$n \sum_{k=r^*+1}^{\min(p,q)} \check{d}_k^2 \rightarrow_d \chi_{(p-r^*)(q-r^*)}^2.$$

Also note that $0 \leq \text{tr}(\mathbf{AB}) \leq \text{tr}(\mathbf{A})\text{tr}(\mathbf{B})$ for any positive semidefinite matrices $\mathbf{A}$ and $\mathbf{B}$ of the same order. It follows that

$$\begin{aligned}
& n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2 \\
&= n \times \text{tr} \left\{ \tilde{\Sigma}_{\varepsilon\varepsilon}^{-1} (\tilde{\mathbf{C}} - \tilde{\mathbf{C}}^{(r)})^\top \tilde{\Sigma}_{xx} (\tilde{\mathbf{C}} \right. \\
&\quad \left. - \tilde{\mathbf{C}}^{(r)}) \leq n \times \text{tr} \left(\tilde{\Sigma}_{\varepsilon\varepsilon}^{-1} \right) \text{tr}(\tilde{\Sigma}_{xx}) \text{tr} \left\{ (\tilde{\mathbf{C}} - \tilde{\mathbf{C}}^{(r)})^\top (\tilde{\mathbf{C}} \right. \\
&\quad \left. - \tilde{\mathbf{C}}^{(r)}) \right\} \\
&= \text{tr} \left(\tilde{\Sigma}_{\varepsilon\varepsilon}^{-1} \right) \text{tr}(\tilde{\Sigma}_{xx}) \\
&\quad \times n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2;
\end{aligned}$$

$$\begin{aligned}
& n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2 \\
&= n \times \text{tr} \left\{ (\tilde{\mathbf{C}} - \tilde{\mathbf{C}}^{(r)})^\top (\tilde{\mathbf{C}} \right. \\
&\quad \left. - \tilde{\mathbf{C}}^{(r)}) \leq n \times \text{tr}(\tilde{\Sigma}_{\varepsilon\varepsilon}) \text{tr} \left(\tilde{\Sigma}_{xx}^{-1} \right) \text{tr} \left\{ \tilde{\Sigma}_{\varepsilon\varepsilon}^{-1} (\tilde{\mathbf{C}} - \tilde{\mathbf{C}}^{(r)})^\top \tilde{\Sigma}_{xx} (\tilde{\mathbf{C}} \right. \\
&\quad \left. - \tilde{\mathbf{C}}^{(r)}) \right\} \\
&= \text{tr}(\tilde{\Sigma}_{xx}) \text{tr} \left(\tilde{\Sigma}_{\varepsilon\varepsilon}^{-1} \right) \\
&\quad \times n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2.
\end{aligned}$$

Therefore,

$$\frac{1}{\text{tr}(\tilde{\Sigma}_{\varepsilon\varepsilon}^{-1}) \text{tr}(\tilde{\Sigma}_{xx})} \times n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2 \leq n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2 \leq \text{tr}(\tilde{\Sigma}_{\varepsilon\varepsilon}) \text{tr} \left(\tilde{\Sigma}_{xx}^{-1} \right) \times n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2,$$

where $\text{tr}(\tilde{\Sigma}_{\varepsilon\varepsilon}^{-1})$, $\text{tr}(\tilde{\Sigma}_{xx})$, $\text{tr}(\tilde{\Sigma}_{\varepsilon\varepsilon})$, and $\text{tr}(\tilde{\Sigma}_{xx}^{-1})$ are all positive and bounded, because $\tilde{\Sigma}_{xx} \rightarrow \Gamma$ and $\tilde{\Sigma}_{\varepsilon\varepsilon} \rightarrow_p \Sigma$ where both Γ and Σ are positive definite. It then follows that

$n \sum_{k=r^*+1}^{\min(p,q)} \tilde{d}_k^2 = O_p(1)$, that is, $n \tilde{d}_k^2 = O_p(1)$ for $k = r^* + 1, \dots, r$.

Proof of Theorem 3.2.1. Recall $\mathbf{U}_{p \times r}^* = (\mathbf{U}^{**}, \mathbf{0})$ and $\mathbf{V}_{q \times r}^* = (\mathbf{V}^{**}, \mathbf{V}^{*\circ})$. For any $p \times r$ matrix $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$ and any $q \times r$ matrix $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$ with $\mathbf{B}_{\mathcal{G}} = \mathbf{0}$, define $\Psi_n(\mathbf{A}, \mathbf{B})$ as follows:

$$\Psi_n(\mathbf{A}, \mathbf{B}) \equiv Q_n\left(\mathbf{U}^* + \frac{1}{\sqrt{n}}\mathbf{A}, \mathbf{V}^* + \frac{1}{\sqrt{n}}\mathbf{B}\right) - Q_n(\mathbf{U}^*, \mathbf{V}^*)$$

$$= -\mathbf{z}_n^\top \text{vec}\left(\frac{1}{\sqrt{n}}\mathbf{X}^\top \mathbf{E}\right) + \frac{1}{2}\mathbf{z}_n^\top \left(\frac{\mathbf{I}_q \otimes \mathbf{X}^\top \mathbf{X}}{n}\right)\mathbf{z}_n + \sum_{k=1}^{r^*} \left\{ \lambda_k^{(n)} \sum_{i,j} w_{ijk} \left(|u_{ik}^* v_{jk}^* + \frac{1}{\sqrt{n}} u_{ik}^* b_{jk} + \frac{1}{\sqrt{n}} a_{ik} v_{jk}^* + \frac{1}{n} a_{ik} b_{jk}| - |u_{ik}^* v_{jk}^*| \right) \right\} + \sum_{k=r^*+1}^r \lambda_k^{(n)}$$

$$\text{where } \mathbf{z}_n = \text{vec}\left(\mathbf{U}^* \mathbf{B}^\top + \mathbf{A} \mathbf{V}^{*\top} + \frac{1}{\sqrt{n}} \mathbf{A} \mathbf{B}^\top\right).$$

Consider the last term in Eq. (11). For any $k(r^* + 1 \leq k \leq r)$, there must exist some $j(1 \leq j \leq q)$ such that $v_{jk}^* \neq 0$. Now suppose for some $i(1 \leq i \leq p)$, $a_{ik} = 0$, then

$$\lambda_k^{(n)} w_{ijl} \left| \frac{1}{\sqrt{n}} a_{ik} v_{jk}^* + \frac{1}{n} a_{ik} b_{jk} \right| = \frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} |\sqrt{n} \tilde{c}_{ijk}|^{-\gamma} |a_{ik} v_{jk}^* + \frac{1}{\sqrt{n}} a_{ik} b_{jk}| \rightarrow \infty$$

as $n \rightarrow \infty$ because $\frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} \rightarrow \infty$ by C4 and $|\sqrt{n} \tilde{c}_{ijk}| = n O_p(d_k) = O_p(1)$ by Lemma 3.2.1. Therefore, this term is bounded only if $a_{ik} = 0$ for any $i = 1, \dots, p$ and $k = r^* + 1, \dots, r$, that is, $\mathbf{A}_2 = \mathbf{0}$. Therefore,

$$\Psi_n(\mathbf{A}, \mathbf{B}) \rightarrow_d \Psi(\mathbf{A}, \mathbf{B}) = \begin{cases} -\mathbf{z}^\top \mathbf{w} + \frac{1}{2} \mathbf{z}^\top (\mathbf{I}_q \otimes \Gamma) \mathbf{z} & \forall i \leq p, k \leq r^*, a_{ik} = 0 \text{ if } u_{ik}^* = 0; \\ & \forall j \leq q, k \leq r^*, b_{jk} = 0 \text{ if } v_{jk}^* = 0; \\ & \forall i \leq p, k = r^* + 1, \dots, r, a_{ik} = 0; \\ \infty & \text{otherwise.} \end{cases}$$

where $\mathbf{z} = \text{vec}(\mathbf{U}^{**} \mathbf{B}_1^\top + \mathbf{A}_1 \mathbf{V}^{**\top})$ and $\mathbf{w} \sim N(\mathbf{0}, \Sigma \otimes \Gamma)$. Note that this limit function no longer involves $\mathbf{V}^\circ$. Similar to Chen et al. (2012), a profile likelihood approach can be adopted to show that $\Psi(\mathbf{A}, \mathbf{B})$ has a unique minimum. The results in (1) can then be established by applying the Argmax theorem (van der Vaart 2000, 81) for $-\Psi_n$. We hence omit the details. Since $\Psi(\mathbf{A}, \mathbf{B})$ attains its minimum only if $a_{ik} = 0$ for any $i = 1, \dots, p$ and $k = r^* + 1, \dots, r$, the result in (2) follows by noting that $a_{ik} = n \hat{u}_{ik} = n O_p(d_k)$. This completes the proof.

Proof of Theorem 3.2.2. Since the singular vectors of the redundant layers are no longer estimable, we write $\hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^\top = \hat{d}_k \hat{\mathbf{u}}_k^\circ \hat{\mathbf{v}}_k^{\circ\top}$ for each $k = r^* + 1, \dots, r$ where $\|\hat{\mathbf{u}}_k^\circ\| = \|\hat{\mathbf{v}}_k^\circ\| = 1$ and focus on studying the estimated singular value $\hat{d}_k$ for each redundant layer.

First, consider $k \in \{1, \dots, r^*\}$. According to the asymptotic normality result stated in

Theorem 3.2.1, $\hat{\mathbf{u}}_{k, \mathcal{A}_k} \rightarrow_p \mathbf{u}_{k, \mathcal{A}_k}^*$ and $\hat{\mathbf{v}}_{k, \mathcal{B}_k} \rightarrow_p \mathbf{v}_{k, \mathcal{B}_k}^*$; thus $\forall i \in \mathcal{A}_k, P(i \in \mathcal{A}_k^{(n)}) \rightarrow 1$, and $\forall j \in \mathcal{B}_k, P(j \in \mathcal{B}_k^{(n)}) \rightarrow 1$. To show (1), it suffices to show that $\forall i \notin \mathcal{A}_k, P(i \in \mathcal{A}_k^{(n)}) \rightarrow 0$, and $\forall j \notin \mathcal{B}_k, P(j \in \mathcal{B}_k^{(n)}) \rightarrow 0$. $\forall j \notin \mathcal{B}_k (v_{jk}^* = 0)$, consider the event $j \in \mathcal{B}_k^{(n)}$. By the KKT optimality conditions,

$$\frac{1}{\sqrt{n}} \mathbf{x}_j^{(v)\top} (\mathbf{y}_k - \mathbf{X}^{(v)} \hat{\mathbf{v}}_k) = \frac{1}{\sqrt{n}} \lambda^{(v)} w_j^{(v)} \quad (12)$$

where $\mathbf{X}^{(v)} = (\mathbf{x}_1^{(v)}, \dots, \mathbf{x}_q^{(v)}) = \mathbf{I}_q \otimes \hat{\mathbf{X}} \mathbf{u}_k$, $\lambda^{(v)} = \lambda_k^{(n)} \sum_{i=1}^p w_i^{(u)} |\hat{u}_{ik}|$, $w_i^{(u)} = |\tilde{u}_{ik}|^{-\gamma}$, $w_j^{(v)} = |\tilde{v}_{jk}|^{-\gamma}$ and $\mathbf{y}_k = \text{vec}(\mathbf{Y} - \mathbf{X} \sum_{h \neq k}^r \hat{\mathbf{u}}_h \hat{\mathbf{v}}_h^\top)$.

Consider the left-hand side of Eq. (12)

$$\begin{aligned} \text{LHS} &= \frac{1}{\sqrt{n}} \mathbf{x}_j^{(v)\top} (\mathbf{y}_k - \mathbf{X}^{(v)} \hat{\mathbf{v}}_k) \\ &= \frac{1}{\sqrt{n}} \left\{ \hat{\mathbf{u}}_k^\top \mathbf{X}^\top \mathbf{X} \left(-\hat{\mathbf{u}}_k \hat{v}_{jk} + \sum_{h \neq k}^{r^*} \mathbf{u}_h^* v_{jh}^* - \sum_{h \neq k}^{r^*} \hat{\mathbf{u}}_h \hat{v}_{jh} - \sum_{h=r^*+1}^r \hat{\mathbf{u}}_h \hat{v}_{jh} \right) + \hat{\mathbf{u}}_k^\top \mathbf{X}^\top \mathbf{e}_{(j)} \right\} \\ &= \hat{\mathbf{u}}_k^\top \frac{\mathbf{X}^\top \mathbf{X}}{n} \left\{ -\hat{\mathbf{u}}_k \sqrt{n} \hat{v}_{jk} + \sqrt{n} \left(\sum_{h \neq k}^{r^*} \mathbf{u}_h^* v_{jh}^* - \sum_{h \neq k}^{r^*} \hat{\mathbf{u}}_h \hat{v}_{jh} \right) - \sqrt{n} \sum_{h=r^*+1}^r \hat{\mathbf{u}}_h \hat{v}_{jh} \right\} + \frac{\hat{\mathbf{u}}_k^\top \mathbf{X}^\top \mathbf{e}_{(j)}}{\sqrt{n}} \\ &= O_p(1). \end{aligned}$$

On the other hand, as $n \rightarrow \infty$, the right-hand side of Eq. (12) is

$$\text{RHS} = \frac{1}{\sqrt{n}} \lambda^{(v)} w_j^{(v)} = \frac{\lambda_k^{(n)}}{\sqrt{n}} |\hat{\mathbf{u}}_k|^\top \mathbf{w}^{(u)} w_j^{(v)} = \frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} |\hat{\mathbf{u}}_k|^\top |\sqrt{n} \tilde{v}_{jk} \hat{\mathbf{u}}_k|^{-\gamma} \rightarrow_p \infty.$$

Therefore, as $n \rightarrow \infty$, $P(j \in \mathcal{B}_k^{(n)}) \leq P\left\{\frac{1}{\sqrt{n}} \mathbf{x}_j^{(v)\top} (\mathbf{y}_k - \mathbf{X}^{(v)} \hat{\mathbf{v}}_k) = \frac{1}{\sqrt{n}} \lambda^{(v)} w_j^{(v)}\right\} \rightarrow 0$.

$\forall i \notin \mathcal{A}_k$, consider the event $i \in \mathcal{A}_k^{(n)}$. By the KKT optimality conditions,

$$\frac{1}{\sqrt{n}} \mathbf{x}_i^{(u)\top} (\mathbf{y}_k - \mathbf{X}^{(u)} \hat{\mathbf{u}}_k) = \frac{1}{\sqrt{n}} \lambda^{(u)} w_i^{(u)} \quad (13)$$

where $\mathbf{X}^{(u)} = (\mathbf{x}_1^{(u)}, \dots, \mathbf{x}_p^{(u)}) = \hat{\mathbf{v}}_k \otimes \mathbf{X}$, $\lambda^{(u)} = \lambda_k^{(n)} \sum_{j=1}^q w_j^{(v)} |\hat{v}_{jk}|$, $w_i^{(u)} = |\tilde{u}_{ik}|^{-\gamma}$,

$w_i^{(v)} = |\tilde{v}_{jk}|^{-\gamma}$ and $\mathbf{y}_k = \text{vec}(\mathbf{Y} - [\mathbf{X} \sum_{h \neq k}^r \hat{\mathbf{u}}_h \hat{\mathbf{v}}_h^\top])$. Consider the left-hand side of Eq. (13)

$$\begin{aligned} \text{LHS} &= \frac{1}{\sqrt{n}} \mathbf{x}_i^{(u)\top} (\mathbf{y}_k - \mathbf{X}^{(u)} \hat{\mathbf{u}}_k) \\ &= \frac{1}{\sqrt{n}} \hat{\mathbf{v}}_k^\top \left\{ \left(\mathbf{v}_k^* \mathbf{u}_k^{*\top} - \hat{\mathbf{v}}_k \hat{\mathbf{u}}_k^\top + \sum_{h \neq k}^{r^*} \mathbf{v}_h^* \mathbf{u}_h^{*\top} - \sum_{h \neq k}^{r^*} \hat{\mathbf{v}}_h \hat{\mathbf{u}}_h^\top - \sum_{h=r^*+1}^r \hat{\mathbf{v}}_h \hat{\mathbf{u}}_h^\top \right) \mathbf{X}^\top \mathbf{x}_{(i)} + \mathbf{E}^\top \mathbf{x}_i \right\} \\ &= \left\{ \sqrt{n} (\hat{v}_k - \mathbf{v}_k^*)^\top \mathbf{v}_k^* \mathbf{u}_k^{*\top} - \hat{\mathbf{v}}_k^\top \hat{\mathbf{v}}_k \sqrt{n} (\hat{\mathbf{u}}_k - \mathbf{u}_k^*)^\top - \sqrt{n} (\hat{\mathbf{v}}_k^\top \hat{\mathbf{v}}_k - \mathbf{v}_k^{*\top} \mathbf{v}_k^*) \mathbf{u}_k^{*\top} \right\} \frac{\mathbf{X}^\top \mathbf{x}_{(i)}}{n} + \hat{\mathbf{v}}_k^\top \left\{ \sqrt{n} \left(\sum_{h \neq k}^{r^*} \mathbf{v}_h^* \mathbf{u}_h^{*\top} - \sum_{h \neq k}^{r^*} \hat{\mathbf{v}}_h \hat{\mathbf{u}}_h^\top \right) - \sqrt{n} \right\} \\ &= O_p(1). \end{aligned}$$

But as $n \rightarrow \infty$, the right-hand side of Eq. (13) is

$$\text{RHS} = \frac{1}{\sqrt{n}} \lambda^{(v)} w_i^{(u)} = \frac{\lambda_k^{(n)}}{\sqrt{n}} |\hat{\mathbf{v}}_k|^\top \mathbf{w}^{(v)} w_i^{(u)} = \frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} |\hat{\mathbf{v}}_k|^\top \sqrt{n} \tilde{u}_{ik} \tilde{v}_k|^{-\gamma} \rightarrow_p \infty.$$

Therefore, as $n \rightarrow \infty$, $P\left(i \in \mathcal{A}_k^{(n)}\right) \leq P\left\{\frac{1}{\sqrt{n}} \mathbf{x}_i^{(u)\top} (\mathbf{y}_k - \mathbf{X}^{(u)} \hat{\mathbf{u}}_k) = \frac{1}{\sqrt{n}} \lambda^{(v)} w_i^{(u)}\right\} \rightarrow 0$.

Now consider $k \in \{r^* + 1, \dots, r\}$. By the KKT condition for d_k^* ,

$$\frac{1}{\sqrt{n}} \text{vec}^\top (\mathbf{X} \hat{\mathbf{u}}_k^\circ \hat{\mathbf{v}}_k^{\circ\top}) [\mathbf{y}_k - \text{vec}(\mathbf{X} \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^\top)] = \frac{1}{\sqrt{n}} \lambda_k^{(n)} \sum_{i=1}^p \sum_{j=1}^q w_{ijk} |\hat{u}_{ik}^\circ \hat{v}_{jk}^\circ| \quad (14)$$

where $\mathbf{y}_k = \text{vec}(\mathbf{Y} - \mathbf{X} \sum_{h \neq k}^r \hat{\mathbf{u}}_h \hat{\mathbf{v}}_h^\top)$. Consider the left-hand side of Eq. (14)

$$\begin{aligned} \text{LHS} &= \frac{1}{\sqrt{n}} \text{vec}^\top (\mathbf{X} \hat{\mathbf{u}}_k^\circ \hat{\mathbf{v}}_k^{\circ\top}) [\mathbf{y}_k - \text{vec}(\mathbf{X} \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^\top)] \\ &= \text{vec}^\top (\hat{\mathbf{u}}_k^\circ \hat{\mathbf{v}}_k^{\circ\top}) \frac{\mathbf{I}_q \otimes \mathbf{X}^\top \mathbf{X}}{n} \left\{ \sqrt{n} \sum_{h=1}^{r^*} \text{vec}(\mathbf{u}_h^* \mathbf{v}_h^{*\top} - \hat{\mathbf{u}}_h \hat{\mathbf{v}}_h^\top) - \sqrt{n} \sum_{h=r^*+1}^r \hat{d}_h \text{vec}(\hat{\mathbf{u}}_h^\circ \hat{\mathbf{v}}_h^{\circ\top}) \right\} + \text{vec}^\top (\hat{\mathbf{u}}_k^\circ \hat{\mathbf{v}}_k^{\circ\top}) \text{vec} \left(\frac{\mathbf{X}^\top \mathbf{E}}{\sqrt{n}} \right) \\ &= O_p(1). \end{aligned}$$

But as $n \rightarrow \infty$, the right-hand side of Eq. (14) is

$$\text{RHS} = \frac{1}{\sqrt{n}} \lambda_k^{(n)} \sum_{i=1}^p \sum_{j=1}^q w_{ijk} |\hat{u}_{ik}^\circ \hat{v}_{jk}^\circ| = \frac{\lambda_k^{(n)}}{\sqrt{n}} n^{\frac{\gamma}{2}} \sum_{i=1}^p \sum_{j=1}^q |\sqrt{n} O_p(\tilde{d}_k)|^{-\gamma} |\hat{u}_{ik}^\circ \hat{v}_{jk}^\circ| \rightarrow_p \infty.$$

Therefore, $P(d_k^* = 0) \rightarrow 0$ for $\forall k \in \{r^* + 1, \dots, r\}$. This completes the proof.

References

- An, H.; Huang, D.; Yao, Q.; Zhang, CH. Technical Report. Department of Statistics, London School of Economics; London, UK: 2008. Stepwise searching for feature variables in high-dimensional linear regression.
- Anderson TW. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*. 1951; 22:327–51.
- Billingsley, P. *Wiley series in probability and statistics: Probability and statistics*. Hoboken, NJ: Wiley; 1999. Convergence of probability measures.
- Bunea F, She Y, Wegkamp M. Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics*. 2011; 39:1282–309.
- Bunea F, She Y, Wegkamp M. Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Annals of Statistics*. 2012; 40:2359–88.
- Bura E, Pfeiffer R. On the distribution of the left singular vectors of a random matrix and its applications. *Statistics and Probability Letters*. 2008; 58:2275–80.
- Chen, K. Dissertation. University of Iowa; Ames, IA: 2011. Regularized multivariate stochastic regression.

- Chen K, Chan KS. Subset arma selection via the adaptive lasso. *Statistics and Its Interface*. 2011; 4:197–205.
- Chen K, Chan KS, Stenseth NC. Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B*. 2012; 74:203–21.
- Chen K, Chan KS, Stenseth NC. Source-sink reconstruction through regularized multicomponent regression analysis—With application to assessing whether North Sea cod larvae contributed to local fjord cod in Skagerrak. *Journal of the American Statistical Association*. 2014; 109:560–73.
- Chen K, Dong H, Chan KS. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*. 2013; 100:901–20. [PubMed: 25045172]
- Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*. 2012; 107:1533–45.
- Efron B, Hastie TJ, Johnstones I, Tibshirani RJ. Least angle regression. *Annals of Statistics*. 2004; 32(2):407–99.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–60.
- Friedman J, Hastie TJ, Höfling H, Tibshirani R. Pathwise coordinate optimization. *Annals of Applied Statistics*. 2007; 2:302–32.
- Hsu PL. On the limit distribution of roots of a determinantal equation. *Journal of London Mathematical Society*. 1941; 16:183–94.
- Huang J, Breheny P, Ma S. A selective review of group selection in high dimensional models. *Statistics Science*. 2012a; 27:81–99.
- Huang J, Wei F, Ma S. Semiparametric regression pursuit. *Statistica Sinica*. 2012b; 22:1403–26. [PubMed: 23559831]
- Izenman AJ. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*. 1975; 5:248–64.
- Knight K, Fu W. Asymptotics for lasso-type estimators. *Annals of Statistics*. 2000; 28:1356–78.
- Lee M, Shen H, Huang JZ, Marron JS. Biclustering via sparse singular value decomposition. *Biometrics*. 2010; 66:1087–95. [PubMed: 20163403]
- Li MC, Chan KS. Multivariate reduced-rank nonlinear time series modeling. *Statistica Sinica*. 2007; 17:139–59.
- Lütkepohl, H. Introduction to multiple time series analysis. New York, NY: Springer Verlag; 1993.
- Ma X, Xiao L, Wong WH. Learning regulatory programs by threshold svd regression. *Proceedings of the National Academy of Sciences of the United States of America*. 2014; 111:15675–80. [PubMed: 25331876]
- Ma, Z.; Sun, T. Adaptive sparse reduced-rank regression. ArXiv e-prints. 2014. <http://arxiv.org/abs/1403.1922>
- Mukherjee A, Zhu J. Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*. 2011; 4:612–22. [PubMed: 22993641]
- Peng J, Zhu J, Bergamaschi A, Han W, Noh DY, Pollack JR, Wang P. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*. 2010; 4:53–77. [PubMed: 24489618]
- R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
- Reinsel, GC.; Velu, P. Multivariate reduced-rank regression: Theory and applications. New York, NY: Springer; 1998.
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–64.
- She Y. Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface*. 2013; 6:197–209.
- Stout WF. The hartman-wintner law of the iterated logarithm for martingales. *Annals of Mathematical Statistics*. 2007; 41:2158–60.
- Tibshirani RJ. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*. 1996; 58:267–88.

- van der Vaart, AW. Asymptotic statistics (Cambridge series in statistical and probabilistic mathematics). Cambridge, UK: Cambridge University Press; 2000.
- Witten DM, Tibshirani RJ, Hastie TJ. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009; 10:515–34. [PubMed: 19377034]
- Yee T, Hastie TJ. Reduced rank vector generalized linear models. *Statistical Modeling*. 2003; 3:367–78.
- Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B*. 2007; 69:329–46.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*. 2006; 68:49–67.
- Zhu H, Khondker Z, Lu Z, Ibrahim JG. Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*. 2014; 109:977–90.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–29.
- Zou H, Hastie TJ. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*. 2005; 67:301–20.
- Zou H, Hastie TJ, Tibshirani RJ. On the degree of freedom of the lasso. *Annals of Statistics*. 2007; 35:2173–92.

Figure 1.

The impact of initial rank specification: (a) the average computation time of RSSVD; (b) average scaled mean-squared error of RSSVD. The SMSEs of RRR when $p = 75$ are also reported as benchmarks. The SNR is set to 0.5.

Figure 2. West German investment, income, and consumption data. The top left panel plots the original data; the other three panels plot the first differences of the logarithm-transformed data.

Table 1

Simulation results for Models I and II, where the SMSEs are reported along with their standard errors in parentheses.

	RSSVD				RRR		RRR OLS	
	SNR	Rank	FDR	FNR	SMSE	Rank	SMSE	SMSE
Model I								
0.5	90.0%	3.1%	0.5%	0.21 (0.2)	100.0%	0.75 (0.4)	4.13 (2.0)	
1	94.7%	1.6%	0.1%	0.09 (0.1)	100.0%	0.39 (0.2)	2.16 (1.1)	
2	96.7%	1.0%	0.0%	0.04 (0.0)	100.0%	0.18 (0.1)	1.04 (0.5)	
Model II								
0.5	86.3%	6.4%	2.9%	0.18 (0.5)	96.3%	2.46 (0.8)	11.6 (5.2)	
1	93.0%	3.9%	0.4%	0.04 (0.2)	98.3%	2.23 (0.7)	7.01 (2.8)	
2	97.7%	3.1%	0.1%	0.02 (0.1)	97.3%	1.98 (0.7)	4.19 (1.4)	

Table 2

Simulation results for Models III and IV, where the SMSEs are reported along with their standard errors in parentheses.

	SNR	RSSVD			RRR		Oracle		OLS
		Rank	FDR	FNR	SMSE	Rank	SMSE	SMSE	
Model I									
	0.5	86.0%	1.6%	17.9%	0.65 (0.4)	94.0%	0.95 (0.5)	0.33 (0.2)	2.75 (1.4)
	1	89.0%	1.1%	9.9%	0.15 (0.1)	99.7%	0.23 (0.1)	0.08 (0.0)	0.66 (0.3)
	2	94.3%	0.2%	5.5%	0.03 (0.0)	100.0%	0.06 (0.0)	0.02 (0.0)	0.16 (0.1)
Model IV									
	0.5	71.3%	1.5%	32.8%	1.91 (1.2)	80.3%	8.63 (2.5)	0.28 (0.2)	17.62 (7.1)
	1	90.7%	0.9%	19.7%	0.50 (0.3)	97.0%	5.68 (1.3)	0.07 (0.0)	7.97 (2.2)
	2	95.0%	0.8%	15.3%	0.25 (0.2)	99.3%	5.11 (1.2)	0.02 (0.0)	5.79 (1.3)