Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 2, *pp.* 203–221

Reduced rank stochastic regression with a sparse singular value decomposition

Kun Chen,

Kansas State University, Manhattan, USA

Kung-Sik Chan

University of Iowa, Iowa City, USA

and Nils Chr. Stenseth

University of Oslo, Norway

[Received August 2010. Final revision July 2011]

Summary. For a reduced rank multivariate stochastic regression model of rank r^*, the regression coefficient matrix can be expressed as a sum of r^* unit rank matrices each of which is proportional to the outer product of the left and right singular vectors. For improving predictive accuracy and facilitating interpretation, it is often desirable that these left and right singular vectors be sparse or enjoy some smoothness property. We propose a regularized reduced rank regression approach for solving this problem. Computation algorithms and regularization parameter selection methods are developed, and the properties of the new method are explored both theoretically and by simulation. In particular, the regularization method proposed is shown to be selection consistent and asymptotically normal and to enjoy the oracle property. We apply the proposed model to perform biclustering analysis with microarray gene expression data.

Keywords: Biclustering; Microarrary gene expression data; Multivariate regression; Oracle property; Regularization

1. Introduction

The reduced rank regression model (Izenman, 1975; Reinsel and Velu, 1998) achieves dimension reduction through restricting the rank of the coefficient matrix. Recently, as high dimensional data become increasingly common, another approach of dimension reduction, through utilizing sparsity-inducing regularization techniques under a multivariate regression framework, has emerged (Turlach *et al.*, 2005; Obozinski *et al.*, 2008; Peng *et al.*, 2010). These two approaches, namely the reduced rank method and regularization technique, both provide very important extensions to classical multivariate regression. Therefore, it is appealing to integrate the two approaches. One novel attempt was made by Yuan *et al.* (2007), in which a regularized least squares approach was proposed to conduct dimension reduction and coefficient estimation simultaneously. The penalty that they considered encourages sparsity in the singular values of the coefficient matrix so that the rank can be automatically reduced and determined as the number of non-zero singular values. However, their model did not take into account possible sparsity structure in the coefficient matrix itself, so it does not do variable selection. Here, we

Address for correspondence: Kung-Sik Chan, Department of Statistics and Actuarial Science, 263 Schaeffer Hall, University of Iowa, Iowa City, IA 52242, USA.
E-mail: kungsik.chan@gmail.com

 1369–7412/12/74203

propose a novel regularized reduced rank regression approach, which concerns the estimation of the sparsity structure in the singular vectors of a reduced rank coefficient matrix. A key advantage of our method is its ability to eliminate both irrelevant responses and predictors and yet to keep the reduced rank structure.

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

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E} \tag{1.1}$$

where $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^{\mathrm{T}}$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^{\mathrm{T}}$, $\mathbf{C}$ is a $p \times q$ coefficient matrix with $\mathrm{rank}(\mathbf{C}) = r^* \leqslant \min(p, q)$ and $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^{\mathrm{T}}$ is a random $n \times q$ matrix; the error vectors are assumed to be independently and identically distributed (IID) with mean vector $E(\mathbf{e}_i) = \mathbf{0}$ and covariance matrix $\mathrm{cov}(\mathbf{e}_i) = \Sigma$, a $q \times q$ positive definite matrix. We assume that the variables are centred so that there is no intercept term. For any $1 \leqslant r \leqslant \min(p, q)$, the rank r least squares estimator of $\mathbf{C}$ which minimizes $\|\mathbf{Y} - \mathbf{XC}\|_{\mathrm{F}}^2$ subject to $\mathrm{rank}(\mathbf{C}) = r$ can be obtained explicitly (Reinsel and Velu, 1998), where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. The rank of $\mathbf{C}$ can also be estimated by various methods; see, for example Anderson (2002, 2003), Camba-Mendez *et al.* (2003) and Yuan *et al.* (2007). Henceforth, the rank of $\mathbf{C}$ is assumed to have been correctly identified to be r^*.

The rank r^* coefficient matrix $\mathbf{C}$ can be expressed through singular value decomposition (SVD) as a sum of r^* unit rank matrices, each of which is proportional to the outer product of the left and right singular vectors, i.e.

$$\mathbf{C} = \mathbf{UDV}^{\mathrm{T}} = \sum_{k=1}^{r^*} d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}} = \sum_{k=1}^{r^*} \mathbf{C}_k, \tag{1.2}$$

where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{r^*})$ consists of r^* orthonormal left singular vectors, $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{r^*})$ consists of r^* orthonormal right singular vectors, $\mathbf{D}$ is an $r^* \times r^*$ diagonal matrix with positive singular values $d_1 > \ldots > d_{r^*}$ on its diagonal and $\mathbf{C}_k = d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$ is the layer k unit rank matrix of $\mathbf{C}$. Here all the singular values are assumed to be distinct so that this SVD decomposition is unique up to the signs of the singular vectors. If some singular values are identical, then the preceding SVD decomposition is non-unique, which complicates the theoretical analysis of the model estimation. In practice, the singular values rarely coincide, so the distinct singular value condition generally holds.

This SVD representation of $\mathbf{C}$ reveals a very appealing latent model interpretation of the reduced rank regression, i.e. $\mathbf{C}$ is composed of r^* orthogonal layers of decreasing importance, and each layer provides a distinct channel or pathway relating the responses to the predictors. In this sense, for each layer k, the elements in $\mathbf{u}_k$ can be viewed as the predictor effects, the elements in $\mathbf{v}_k$ can be viewed as the response effects and the singular value d_k indicates the relative importance of the association. Under this reduced rank structure, a more practical situation is that each channel or pathway of association between the responses and predictors may involve only a subset of the responses and predictors. Therefore, to achieve further dimension reduction and to facilitate interpretation, it is desirable that the left and right singular vectors be sparse.

For example, in a microarray biclustering problem (Busygin *et al.*, 2008; Lee *et al.*, 2010), the data $\mathbf{Y}$ consist of expression levels of thousands of genes, measured from a few subjects, who are either normal subjects or patients with different types of lung cancer. The goal is to identify sets of biologically relevant genes that are expressed at different levels in different types of cancer. A biclustering analysis can be performed by seeking a sparse SVD approximation of the matrix $\mathbf{Y}$, which corresponds to model (1.1) with $\mathbf{X}$ being an identity matrix. Our method may be heuristically justified as follows:

(a) the fact that the subjects form a few cancer groups indicates that the cancer–gene associations admit a reduced rank structure and
(b) that each cancer–gene association generally involves certain subsets of genes and subjects implies a low rank model with sparse components.

Model (1.1) with a non-identity design matrix may, however, be used to facilitate supervised learning. For instance, if the types of cancer of the subjects are known, supervised gene clustering may be performed by fitting model (1.1) with $\mathbf{X}$ incorporating the cancer type information; see Section 4.4 for details. Another example is an ecological application, in which we analyse a data set quantifying the abundance of Norwegian Skagerrak coastal cod, for simultaneously capturing the North Sea cod spawning window and identifying the set of coastal fjords that are influenced by the larval drift from the North Sea (Stenseth *et al.*, 2006). The fact that there are only a few cod spawning populations in the North Sea suggests that a reduced rank model is appropriate. It is hypothesized that, among the 18 coastal fjords under consideration, only those exposed to the Skagerrak could potentially receive larval drift from the North Sea; hence the right singular vector of the coefficient matrix, which comprises the fjord-specific larval drift effects, is expected to be sparse. However, it is believed that the spawning window lasts for only 2 weeks somewhere during a 45-day period under study, so the left singular vector which consists of the daily spawning effects is believed to be smooth in time and hump shaped. Therefore, to address these problems under the multivariate regression framework, the main challenge is how to recover certain sparsity or smoothness structure in a reduced rank regression; details of this ecological application will be published elsewhere. Similar data structures also arise in many other problems in various fields including genomewide association studies (Vounou *et al.*, 2010).

To recover an SVD structure that is both sparse and orthogonal is certainly not an easy task. However, strict orthogonality is often not necessary in real applications (Lee *et al.*, 2010). We thus consider a local search strategy which relaxes the orthogonality condition between the estimated left or right singular vectors to accommodate efficient search for their sparsity patterns. We propose to estimate $\mathbf{C}$ by minimizing the following objective function with respect to the triplets $(d_k, \mathbf{u}_k, \mathbf{v}_k)$ for $k = 1, \ldots, r^*$:

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\sum_{k=1}^{r^*} d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}\|_{\mathrm{F}}^2 + \sum_{k=1}^{r^*} \mathrm{Pe}\{\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)\}, \tag{1.3}$$

where $\|\mathbf{u}_k\| = \|\mathbf{v}_k\| = 1$ with $\|\cdot\|$ denoting the l_2-norm, $\Sigma_{k=1}^{r^*} d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$ corresponds to the SVD decomposition of $\mathbf{C}$ but strict orthogonality among $\mathbf{u}_k$s and $\mathbf{v}_k$s is not required (more details are given later), $\mathrm{Pe}(\cdot)$ is some penalty function and λ_ks are the regularization parameters controlling the degrees of penalization of distinct layers. To prompt sparsity in each pair of $\mathbf{u}_k$ and $\mathbf{v}_k$, we consider

$$\begin{aligned}\mathrm{Pe}\{\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)\} &= \lambda_k \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ijk}|d_k u_{ik} v_{jk}| \\ &= \lambda_k (w_k^{(d)} d_k)\left(\sum_{i=1}^{p} w_{ik}^{(u)}|u_{ik}|\right)\left(\sum_{j=1}^{q} w_{jk}^{(v)}|v_{jk}|\right),\end{aligned} \tag{1.4}$$

where $w_{ijk} = w_k^{(d)} w_{ik}^{(u)} w_{jk}^{(v)}$ are possibly data-driven weights to be elaborated below. This penalty term is analogous to the adaptive lasso penalty (Zou, 2006), i.e. it is proportional to the weighted l_1-norm of an SVD layer $d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$. From this point of view, it assigns the correct amount of penalization to each element of the SVD layer. Interestingly, owing to its multiplicative form as shown in equation (1.4), it can also be viewed as penalizing each of the singular vectors comprising the SVD layer, which leads to automatic adjustment of the possibly different degrees of sparsity

between $\mathbf{u}_k$ and $\mathbf{v}_k$. Another advantage of the penalty term in equation (1.4) is that only one regularization parameter is required for each pair of singular vectors.

The objective function (1.3) is non-convex and involves multiple regularization parameters. We have developed an efficient optimization algorithm which can be implemented in parallel. Moreover, our sparse SVD estimator is shown to enjoy the oracle properties, i.e. it recovers the correct sparse SVD structure with probability tending to 1 as the sample size goes to ∞.

The rest of the paper is organized as follows. We develop the methodology for the unit rank case in Section 2, with the extension to the higher rank cases elaborated in Section 3. The biclustering application and several simulation studies illustrating the new methods are given in Section 4. Some asymptotic results of the method proposed are presented in Section 5. We conclude in Section 6.

2. Sparse unit rank regression

2.1. Optimization algorithm and initial values

When $r^* = 1$, the problem reduces to minimizing the following objective function with respect to the triplets $(d, \mathbf{u}, \mathbf{v})$:

$$\tfrac{1}{2}\|\mathbf{Y} - d\mathbf{X}\mathbf{u}\mathbf{v}^{\mathrm{T}}\|_{\mathrm{F}}^2 + \lambda \sum_{i=1}^{p}\sum_{j=1}^{q} w_{ij}|du_i v_j|, \tag{2.1}$$

where $d\mathbf{u}\mathbf{v}^{\mathrm{T}}$ is the SVD of the unit rank coefficient matrix $\mathbf{C}$, i.e. $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$, and $w_{ij} = w^{(d)}w_i^{(u)}w_j^{(v)}$ are data-driven weights. Note that the layer subscript 'k' is dropped from all the notation for simplicity.

Assume that some $\sqrt{n}$-consistent estimator $\tilde{\mathbf{C}}$ of $\mathbf{C}$ is available, e.g. the reduced rank least squares estimator (Reinsel and Velu, 1998), whose SVD is given by $\tilde{d}\tilde{\mathbf{u}}\tilde{\mathbf{v}}^{\mathrm{T}}$. Using the perturbation expansion of matrices (theorem 3, Izenman (1975)), it can be readily shown that the $\sqrt{n}$-consistency of $\tilde{\mathbf{C}}$ implies that $\tilde{d}$, $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are all $\sqrt{n}$-consistent estimators of d, $\mathbf{u}$ and $\mathbf{v}$ respectively. Following Zou (2006), the weights can be chosen as

$$\left.\begin{aligned} w^{(d)} &= |\tilde{d}|^{-\gamma}, \\ \mathbf{w}^{(u)} &= (w_1^{(u)}, \ldots, w_p^{(u)})^{\mathrm{T}} = |\tilde{\mathbf{u}}|^{-\gamma}, \\ \mathbf{w}^{(v)} &= (w_1^{(v)}, \ldots, w_q^{(v)})^{\mathrm{T}} = |\tilde{\mathbf{v}}|^{-\gamma}, \end{aligned}\right\} \tag{2.2}$$

where γ is a prespecified non-negative parameter and $|\cdot|^{-\gamma}$ is defined componentwise for the enclosed vector. Here, on the basis of extensive simulations and as suggested by Zou (2006), we use $\gamma = 2$ in all numerical studies.

The objective function (2.1) admits a multiconvex structure (Gorski *et al.*, 2007). For fixed $\mathbf{u}$, minimization of function (2.1) with respect to $(d, \mathbf{v})$ becomes minimization with respect to $\check{\mathbf{v}} = \mathrm{diag}(d\mathbf{w}^{(v)})\mathbf{v}$ of

$$\tfrac{1}{2}\|\mathbf{y} - \mathbf{X}^{(v)}\check{\mathbf{v}}\|^2 + \lambda^{(v)} \sum_{j=1}^{q} |\check{v}_j|, \tag{2.3}$$

where $\mathrm{diag}(\mathbf{a})$ denotes a diagonal matrix with entries of $\mathbf{a}$ on its diagonal, $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\mathbf{X}^{(v)} = \mathrm{diag}(\mathbf{w}^{(v)})^{-1} \otimes (\mathbf{X}\mathbf{u})$ and $\lambda^{(v)} = \lambda w^{(d)}(\Sigma_{i=1}^{p} w_i^{(u)}|u_i|)$. The symbol '$\otimes$' denotes the Kronecker product and we shall freely make use of several properties of the Kronecker product; see Schott (2005), chapter 8. Model (2.3) can be recognized as a lasso regression problem with respect to $\check{\mathbf{v}}$, without an intercept term. Moreover, note that $\mathbf{X}^{(v)}$ is always an orthogonal matrix; hence the solution of problem (2.3) is explicit (Tibshirani, 1996); see lemma 1 below. In contrast, for fixed $\mathbf{v}$,

minimization of function (2.1) with respect to $(d, \mathbf{u})$ becomes minimization with respect to $\check{\mathbf{u}} = \mathrm{diag}(d\mathbf{w}^{(u)})\mathbf{u}$ of

$$\tfrac{1}{2}\|\mathbf{y} - \mathbf{X}^{(u)}\check{\mathbf{u}}\|^2 + \lambda^{(u)} \sum_{i=1}^{p} |\check{u}_i|, \tag{2.4}$$

where $\mathbf{X}^{(u)} = \mathbf{v} \otimes \mathbf{X}\,\mathrm{diag}(\mathbf{w}^{(u)})^{-1}$ and $\lambda^{(u)} = \lambda w^{(d)}(\Sigma_{j=1}^{q} w_j^{(v)}|v_j|)$. Again, this is a lasso regression problem with respect to $\check{\mathbf{u}}$, without an intercept term. However, the design matrix $\mathbf{X}^{(u)}$ is orthogonal if and only if $\mathbf{X}$ is orthogonal.

We can take advantage of the multiconvex structure of the objective function (2.1) in optimization. Here are the steps of our numerical *sparse unit rank regression algorithm* for a fixed λ.

(a) Choose a non-zero initial value for $\hat{\mathbf{u}}$.
(b) Given $\mathbf{u} = \hat{\mathbf{u}}$, minimize function (2.3) to obtain $\check{\mathbf{v}}$. Let $\hat{d} = \|\mathrm{diag}(\mathbf{w}^{(v)})^{-1}\check{\mathbf{v}}\|$ and $\hat{\mathbf{v}} = \mathrm{diag}(\hat{d}\mathbf{w}^{(v)})^{-1}\check{\mathbf{v}}$.
(c) Given $\mathbf{v} = \hat{\mathbf{v}}$, minimize function (2.4) to obtain $\check{\mathbf{u}}$. Let $\hat{d} = \|\mathrm{diag}(\mathbf{w}^{(u)})^{-1}\check{\mathbf{u}}\|$ and $\hat{\mathbf{u}} = \mathrm{diag}(\hat{d}\mathbf{w}^{(u)})^{-1}\check{\mathbf{u}}$.
(d) Repeat steps (b) and (c), until $\hat{\mathbf{C}} = \hat{d}\hat{\mathbf{u}}\hat{\mathbf{v}}^{\mathrm{T}}$ converges, i.e. $\|\hat{\mathbf{C}}_{\mathrm{c}} - \hat{\mathbf{C}}_{\mathrm{p}}\|_{\mathrm{F}}/\|\hat{\mathbf{C}}_{\mathrm{p}}\|_{\mathrm{F}} < \varepsilon$, where $\hat{\mathbf{C}}_{\mathrm{c}}$ is the current fit, $\hat{\mathbf{C}}_{\mathrm{p}}$ is the previous fit and ε is the level of tolerance, e.g. $\varepsilon = 10^{-6}$.

This algorithm could also start from updating $\mathbf{u}$ with steps (b) and (c) reversed. In either case, the algorithm uses a block co-ordinate descent structure with two overlapping blocks of parameters, i.e. $(d, \mathbf{u})$ and $(d, \mathbf{v})$. Within each block, the model is transformed to a lasso regression problem, for which several fast algorithms have been developed, e.g. the algorithm LARS (Efron *et al.*, 2004) and the co-ordinate descent algorithm (Friedman *et al.*, 2007). It is clear that the objective function decreases monotonically along the iterations. The algorithm is therefore stable and guaranteed to converge, although not necessarily to the global minimum of the objective function. A multiconvex optimization problem may have multiple local minima, and its non-convex optimization requires more complicated algorithms (Gorski *et al.*, 2007). Nevertheless, our limited experience suggests that the iterative algorithm proposed works well.

The following lemma concerns the orthogonal design case for a fixed λ. It shows that, conditional on either $\mathbf{u}$ or $\mathbf{v}$, the other one along with the singular value d can be estimated via a simple soft threshholding rule (Donoho and Johnstone, 1995), which is due to the fact that the solution of a lasso problem is explicit under orthogonal design (Tibshirani, 1996).

Lemma 1. Suppose that $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \Lambda$, where Λ is a diagonal matrix with positive diagonal elements. Then the solution $(\hat{d}, \hat{\mathbf{u}}, \hat{\mathbf{v}})$ of model (2.1) satisfies the equations

$$\hat{d}\hat{\mathbf{u}} = \mathrm{sgn}(\mathbf{X}^{\mathrm{T}}\mathbf{Y}\hat{\mathbf{v}}) \circ \{\Lambda^{-1}(|\mathbf{X}^{\mathrm{T}}\mathbf{Y}\hat{\mathbf{v}}| - \lambda^{(u)}\mathbf{w}^{(u)})_+\},$$

$$\hat{d}\hat{\mathbf{v}} = \mathrm{sgn}(\mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{u}}) \circ \left\{\frac{1}{\hat{\mathbf{u}}^{\mathrm{T}}\Lambda\hat{\mathbf{u}}}(|\mathbf{Y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{u}}| - \lambda^{(v)}\mathbf{w}^{(v)})_+\right\},$$

where $\lambda^{(u)} = \lambda w^{(d)}\Sigma_{j=1}^{q} w_j^{(v)}|\hat{v}_j|$, $\lambda^{(v)} = \lambda w^{(d)}\Sigma_{i=1}^{p} w_i^{(u)}|\hat{u}_i|$, and the symbol '$\circ$' denotes the Hadamard product, which is also known as the entrywise product.

The estimated coefficients vary with λ and produce a path of solutions regularized by λ. In practice, the relevant range of λ equals $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\max}$ is the smallest λ at which all penalized coefficients are 0, and $\lambda_{\min}$ is either 0 or a minimum value at which the model becomes excessively large in terms of number of non-zero parameters or model estimation becomes numerically unstable (Breheny and Huang, 2009). To find the solution path, we suggest starting at $\lambda_{\max} - \varepsilon$, where ε is a small positive number, and proceeding towards $\lambda_{\min}$. Because the path is continuous, the estimate from the previous value of λ can be used as the

initial value for the next value of λ to speed up the computation. This approach works very well in practice, and the algorithm usually converges within only a few iterations. We also point out that the reverse approach, which starts from small λ and goes towards $\lambda_{\max}$, may fail occasionally. This is because when λ proceeds close to $\lambda_{\max}$ from below, with an overfitted initial value, the inner updating step could produce zero solution for either the left or right singular vector, so the algorithm cannot proceed. The following lemma determines $\lambda_{\max}$ and the initial non-zero solution of problem (2.1) corresponding to $\lambda_{\max} - \varepsilon$ explicitly.

Lemma 2. Denote $\mathbf{Y} = (\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(q)})$ and $\mathbf{X} = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(p)})$. Then

$$\lambda_{\max} = \max\left\{ \left| \frac{1}{w_{ij}} \mathbf{x}_{(i)}^{\mathrm{T}} \mathbf{y}_{(j)} \right|, i = 1, \ldots, p; j = 1, \ldots, q \right\},$$

where $w_{ij} = w^{(d)} w_i^{(u)} w_j^{(v)}$. Moreover, letting $(i^*, j^*) = \arg\max_{(i,j)} |(1/w_{ij}) \mathbf{x}_{(i)}^{\mathrm{T}} \mathbf{y}_{(j)}|$, then the initial non-zero singular vectors of problem (2.1) denoted as $(\mathbf{u}^{(0)}, \mathbf{v}^{(0)})$ are given by

$$u_{i^*}^{(0)} = 1, \qquad u_i^{(0)} = 0, \forall i = 1, \ldots, p \text{ and } i \neq i^*,$$
$$v_{j^*}^{(0)} = \mathrm{sgn}(\mathbf{x}_{(i^*)}^{\mathrm{T}} \mathbf{y}_{(j^*)}), \qquad v_j^{(0)} = 0, \forall j = 1, \ldots, q \text{ and } j \neq j^*.$$

Proof. Note that the minimization problem (2.1) has the same $\lambda_{\max}$ as the lasso model $\frac{1}{2}\|\mathbf{y} - \mathbf{H}\rho\|^2 + \lambda \sum_{i=1}^{pq} |\rho_i|$, where $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\mathbf{H} = \mathrm{diag}(w^{(d)} \mathbf{w}^{(v)})^{-1} \otimes (\mathbf{X}\, \mathrm{diag}(\mathbf{w}^{(u)})^{-1})$ and $\rho = (\rho_1, \ldots, \rho_{pq})^{\mathrm{T}}$ is a $pq \times 1$ vector. Then $\lambda_{\max}$ and the initial non-zero solution can be obtained as above by the Karush–Kuhn–Tucker optimality conditions for the lasso problem.

We have implemented our algorithms in R (R Development Core Team, 2008). For all the numerical studies, we follow the approach of Friedman *et al.* (2010) and compute solutions along a grid of 100 λ-values that are equally spaced on the log-scale.

2.2. Regularization parameter selection

Once a regularization path has been fitted, it is important to be able to choose an optimal point along the path. For small-scale problems, the optimal λ can be chosen by K-fold cross-validation (CV), based on the predictive performance of the models (Stone, 1974). For large-scale problems, the CV method can be computationally expensive; hence alternative approaches including generalized CV and information criteria have been widely used. Although all these methods may be applicable here, we propose a Bayesian information criterion BIC because of its computational efficiency and promising performances on variable selection.

Denote $(\hat{d}^{(\lambda)}, \hat{\mathbf{u}}^{(\lambda)}, \hat{\mathbf{v}}^{(\lambda)})$ as the fitted value of $(d, \mathbf{u}, \mathbf{v})$ with the regularization parameter being λ, and define BIC as

$$\mathrm{BIC}(\lambda) = \log\{\mathrm{SSE}(\lambda)\} + \frac{\log(qn)}{qn} \mathrm{df}(\lambda), \tag{2.5}$$

where $\mathrm{SSE}(\lambda) = \|\mathbf{Y} - \hat{d}^{(\lambda)} \mathbf{X} \hat{\mathbf{u}}^{(\lambda)} \hat{\mathbf{v}}^{(\lambda)\mathrm{T}}\|_{\mathrm{F}}^2$ denotes the sum of squared error, and $\mathrm{df}(\lambda)$ is the effective number of parameters or the degrees of freedom of the model.

Zou *et al.* (2007) showed that the number of non-zero coefficients is an unbiased estimator of the degrees of freedom for the lasso problem. The non-convex objective function (2.1) admits a conditional lasso structure, as shown in the preceding section. Therefore, we propose the following estimator for $\mathrm{df}(\lambda)$:

$$\widehat{\mathrm{df}}(\lambda) = \sum_{i=1}^{p} I(\hat{u}_i^{(\lambda)} \neq 0) + \sum_{j=1}^{q} I(\hat{v}_j^{(\lambda)} \neq 0) - 1,$$

where $I(\cdot)$ is the indicator function. Note that 1 degree of freedom is lost because there are two unitary constraints ($\|\mathbf{u}\| = 1$ and $\|\mathbf{v}\| = 1$) and one additional free parameter d. We examine the performance of the proposed criterion via simulation in Section 4.

3. Extension to the higher rank cases

To obtain sparse estimates of multiple layers, one naive approach consists of sequentially performing the proposed sparse unit rank regression, each time with the data matrix $\mathbf{Y}$ replaced by the residual matrix that is obtained by subtracting previously estimated layers from the original data matrix. We refer to this method as the *sequential extraction algorithm* (SEA). This idea has been used in many penalized matrix decomposition problems (Witten *et al.*, 2009; Lee *et al.*, 2010) in which $\mathbf{X}$ is an identity matrix, and the rationale is that, for an unpenalized model ($\lambda = 0$), the SEA method can sequentially extract the SVD layers of the data matrix $\mathbf{Y}$ itself. However, in general regression settings, the sequentially extracted layers by the SEA can be shown to correspond to another interesting decomposition of the coefficient matrix (note that SVD is only one of many decompositions of a matrix), and it need not produce SVD layers of the coefficient matrix $\mathbf{C}$, so it is not suitable for recovering the desired sparse SVD structure in $\mathbf{C}$; see the details in the supplementary materials.

Here we propose an *exclusive extraction algorithm* (EEA). The idea is to seek a $\hat{\mathbf{C}}$ with sparse SVD structure near some initial consistent estimator, e.g. the least squares reduced rank regression estimator $\tilde{\mathbf{C}}$, whose SVD is given by $\Sigma_{k=1}^{r^*} \tilde{d}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^{\mathrm{T}} = \Sigma_{k=1}^{r^*} \tilde{\mathbf{C}}_k$. The problem can then be decomposed into r^* parallel sparse unit rank regression problems, by forming r^* 'exclusive layers' $\mathbf{Y}_k$ ($k = 1, \ldots, r^*$) based on $\tilde{\mathbf{C}}$. The EEA is as follows.

(a) For each $k \in \{1, \ldots, r^*\}$:
   (i) construct the adaptive weights $w_k^{(d)} = |\tilde{d}_k|^{-\gamma}$, $\mathbf{w}_k^{(u)} = |\tilde{\mathbf{u}}_k|^{-\gamma}$ and $\mathbf{w}_k^{(v)} = |\tilde{\mathbf{v}}_k|^{-\gamma}$;
   (ii) construct the exclusive layer $\mathbf{Y}_k = \mathbf{Y} - \mathbf{X}(\tilde{\mathbf{C}} - \tilde{\mathbf{C}}_k)$;
   (iii) find $(\hat{d}_k, \hat{\mathbf{u}}_k, \hat{\mathbf{v}}_k)$ by performing the sparse unit rank regression of $\mathbf{Y}_k$ on $\mathbf{X}$. The optimal λ_k can be chosen by either CV or some information criterion, e.g. BIC.

(b) The final estimator of $\mathbf{C}$ is given by $\hat{\mathbf{C}} = \Sigma_{k=1}^{r^*} \hat{d}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^{\mathrm{T}}$.

In the above EEA algorithm, the computational cost increases linearly in r^*, and the estimation for different layers can be performed in parallel. Our simulation studies confirm that this one-step algorithm works well. Moreover, with suitable choice of λ_k, this one-step EEA estimator is $\sqrt{n}$ consistent and it also consistently estimates the sparsity pattern in the singular vectors; see the details in the supplementary materials.

In practice, the quality of the EEA estimation may partly depend on the initial estimator of $\mathbf{C}$ which is used to form the exclusive layers. However, since the EEA estimator is consistent, it can be used to form presumably more accurate exclusive layers. Therefore, to improve estimation, the EEA algorithm can be performed iteratively, each time using the previous sparse estimates as initial values to refine the estimation. We call this method the *iterative exclusive extraction algorithm* (IEEA).

(a) Start from some initial estimator $\hat{\mathbf{C}}^{(0)}$ (e.g. $\tilde{\mathbf{C}}$) to form the exclusive layers and to construct adaptive weights; perform the EEA method to obtain $\hat{\mathbf{C}}^{(1)}$.
(b) On the basis of $\hat{\mathbf{C}}^{(i)}$, perform the EEA method to obtain $\hat{\mathbf{C}}^{(i+1)}$.
(c) Repeat step (b) until $\hat{\mathbf{C}}^{(i)}$ converges according to some stopping criterion, e.g. $\|\hat{\mathbf{C}}^{(i+1)} - \hat{\mathbf{C}}^{(i)}\| / \|\hat{\mathbf{C}}^{(i)}\| < \varepsilon$, where ε is the level of tolerance, e.g. $\varepsilon = 10^{-6}$.

In all simulation studies and real applications that are reported below, we used the reduced rank least squares estimator as the initial estimator for both the EEA and IEEA algorithms. For $p > n$, the initial least squares estimator cannot be computed because the Gram matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is singular, in which case a small positive constant, e.g. 10^{-4}, can be added to the diagonal elements to make the matrix invertible, which was done in our numerical studies. We note that it has recently been shown by Bunea *et al.* (2011) that the consistency result of the reduced rank least squares estimator (using the Moore–Penrose inverse) can be extended to high dimensional situations, i.e. $p > n$. Alternative initial estimators are the ridge regression estimator and the nuclear norm penalized least squares estimator (Yuan *et al.*, 2007).

The IEEA algorithm can be regarded as a way to solve the general optimization problem (1.3), i.e. the algorithm has a block co-ordinate descent structure, with each SVD layer as one block. To see this clearly, suppose that the current estimate is $\hat{\mathbf{C}}^{(i)} = \Sigma_{k=1}^{r^*} \hat{\mathbf{C}}_k^{(i)}$. Then, for each $k = 1, \ldots, r^*$, the objective function (1.3) conditional on $\hat{\mathbf{C}}_r^{(i)}$, $r \in \{1, \ldots, r^*\}$ and $r \neq k$, can be written as $\frac{1}{2}\|\mathbf{Y}_k^{(i)} - d_k \mathbf{X}\mathbf{u}_k\mathbf{v}_k^{\mathrm{T}}\|_{\mathrm{F}}^2 + \mathrm{Pe}\{\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)\} + \text{constant}$, where $\mathbf{Y}_k^{(i)} = \mathbf{Y} - \mathbf{X}(\hat{\mathbf{C}}^{(i)} - \hat{\mathbf{C}}_k^{(i)})$, which is exactly the objective function that we solve to obtain $\hat{\mathbf{C}}_k^{(i+1)}$ in the IEEA algorithm. The selection of the regularization parameters is nested within the iterative algorithm, which avoids using the computationally expensive grid search of r^* regularization parameters. Our experience from simulation studies and real applications suggests that the estimation can be substantially improved on only one or two additional iterations, and the IEEA algorithm typically converges within only a few iterations.

The optimization is carried out locally near an initial consistent estimator of $\mathbf{C}$ and, to accommodate efficient search of the sparsity pattern within each of the singular vectors, the exact orthogonality among the left or right singular vectors is not enforced. Consequently, the algorithms that are presented here may not produce exact orthogonality among the estimated layers. However, we have shown that our estimators of different layers are consistent, and the relaxation of exact orthogonality improves local search efficiency and yet preserves the oracle properties; see Section 5 and the on-line supplementary materials. So in this sense our sparse estimators of the SVD layers enjoy asymptotic orthogonality. Although it remains an open problem to derive an efficient estimation method for finding jointly sparse and orthogonal SVD layers, we believe that our proposed methods suffice for most applications.

4. Simulation and real applications

4.1. Unit rank biclustering

In this simulation study, we consider a unit rank biclustering problem. Let $\mathbf{C} = d\mathbf{u}\mathbf{v}^{\mathrm{T}}$ be a 50×100 unit rank matrix ($p = 50$ and $q = 100$) with $d = 50$ and

$$\check{\mathbf{u}} = (10, -10, 8, -8, 5, -5, \mathrm{rep}(3, 5), \mathrm{rep}(-3, 5), \mathrm{rep}(0, 34))^{\mathrm{T}},$$
$$\mathbf{u} = \check{\mathbf{u}}/\|\check{\mathbf{u}}\|,$$
$$\check{\mathbf{v}} = (10, 9, 8, 7, 6, 5, 4, 3, \mathrm{rep}(2, 17), \mathrm{rep}(0, 75))^{\mathrm{T}},$$
$$\mathbf{v} = \check{\mathbf{v}}/\|\check{\mathbf{v}}\|,$$

where $\mathrm{rep}(a, b)$ denotes a vector of length b, whose entries are all equal to a. Let $\mathbf{Y} = \mathbf{C} + \mathbf{E}$, where the elements of $\mathbf{E}$ are IID samples from $N(0, 1)$, which makes the signal-to-noise ratio SNR approximately equal to 0.5.

Lee *et al.* (2010) used this example to compare their sparse SVD (SSVD) method with several other popular biclustering methods. Their model corresponds to the special case of model (1.1) that $\mathbf{X}$ is a 50×50 identity matrix ($p = n = 50$) with an additive penalty form $\lambda^{(u)} \Sigma_{i=1}^{p} w_i^{(u)} |du_i| +$

Table 1. Comparison of sparse reduced rank regression (SRRR) with the SSVD method proposed by Lee *et al.* (2010)

Method		*Average and % of correctly identified 0s*	*Average and % of correctly identified non-0s*	*Misclassification rate (%)*
SRRR	**v**	73.69 (98.25%)	24.78 (99.13%)	1.53
	u	33.90 (99.70%)	16.00 (100.0%)	0.21
	Overall	107.58 (98.70%)	40.78 (99.46%)	1.09
SSVD	**v**	73.95 (98.60%)	24.73 (98.93%)	1.32
	u	33.77 (99.32%)	16.00 (100.0%)	0.46
	Overall	107.71 (98.82%)	40.73 (99.35%)	1.04

$\lambda^{(v)}\Sigma_{j=1}^{q} w_j^{(v)}|dv_j|$, where $\lambda^{(u)}$ and $\lambda^{(v)}$ are two distinct regularization parameters. The SSVD method is effected via block co-ordinate descent, i.e. alternately updating $\mathbf{u}$ and $\mathbf{v}$, with $\lambda^{(u)}$ and $\lambda^{(v)}$ also alternately selected by BIC within the co-ordinate descent iterations (Lee *et al.*, 2010). Therefore, the SSVD method can be regarded as a special case of our proposed sparse reduced rank regression method, although Lee *et al.* (2010) did not make a connection with the reduced rank regression and did not study the theoretical properties of their estimator. Nevertheless, the SSVD method was found to outperform other methods greatly in terms of recovering the desired sparsity structure. Hence, we compare our method with only the SSVD method.

We use our proposed sparse unit rank regression method (under orthogonal design), and the optimal λ is chosen on the basis of BIC. We replicated the experiment 1000 times. Table 1 contrasts the simulation results of our method with those of the SSVD method. It shows that our method enjoys extremely low misclassification rates, which are comparable with those of the SSVD method; the misclassification rate of identifying $\mathbf{v}$, for example, is defined as the average proportion of both incorrectly identified 0s and non-zeros in $\hat{\mathbf{v}}$ from all runs. We have also compared the two methods in terms of estimation accuracy which is measured by the average scaled mean-squared error SMSE from all runs, i.e. $\mathrm{SMSE} = 100\|\mathbf{C} - \hat{\mathbf{C}}\|_{\mathrm{F}}^2/pq$, and they are very similar to each other (sparse reduced rank regression, 1.33; SSVD, 1.27). Note that only one regularization parameter is used in our method, but the more parsimonious penalty term (1.4) is capable of recovering the correct amount of sparsity in both $\mathbf{u}$ and $\mathbf{v}$. The computation of our method is very fast, and the whole simulation exercise took only a few seconds.

4.2. Higher rank examples

We compare the performances on sparse SVD recovery of the SEA and IEEA estimators and compare their prediction and estimation performances with the ordinary least squares (OLS) estimator, reduced rank regression (RRR) estimator and the nuclear norm penalized (NNP) least squares estimator that was proposed by Yuan *et al.* (2007) in the general regression setting. We construct $\mathbf{X}$ by generating its n rows as IID samples from a multivariate normal distribution $\mathrm{MVN}(\mathbf{0}, \Gamma)$, where $\Gamma = (\Gamma_{ij})_{p\times p}$ and $\Gamma_{ij} = 0.5^{|i-j|}$. Two scenarios in terms of moderate *versus* high model dimensions are considered, i.e. model I, $p, q < n$, and model II, $p, q > n$.

(a) Model I ($p = q = 25, n = 50$ and $r^* = 3$): let $\mathbf{C}$ be a 25×25 rank 3 matrix whose SVD is given by $\Sigma_{k=1}^{3} d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$ with $d_1 = 20, d_2 = 10$ and $d_3 = 5$. The $\mathbf{u}_k$s are given by

$$\check{\mathbf{u}}_1 = (\text{unif}(\mathcal{A}_u, 5), \text{rep}(0, 20))^{\mathrm{T}},$$
$$\check{\mathbf{u}}_2 = (\text{rep}(0, 5), \text{unif}(\mathcal{A}_u, 5), \text{rep}(0, 15))^{\mathrm{T}},$$
$$\check{\mathbf{u}}_3 = (\text{rep}(0, 10), \text{unif}(\mathcal{A}_u, 5), \text{rep}(0, 10))^{\mathrm{T}},$$
$$\mathbf{u}_k = \check{\mathbf{u}}_k / \|\check{\mathbf{u}}_k\| \text{ for } k = 1, 2, 3,$$

and the $\mathbf{v}_k$s are given by

$$\check{\mathbf{v}}_1 = (\text{unif}(\mathcal{A}_v, 10), \text{rep}(0, 15))^{\mathrm{T}},$$
$$\check{\mathbf{v}}_2 = (\text{rep}(0, 12), \text{unif}(\mathcal{A}_v, 10), \text{rep}(0, 3))^{\mathrm{T}},$$
$$\check{\mathbf{v}}_3 = (\text{rep}(0, 6), \check{\mathbf{v}}_{1,7:8}, -\check{\mathbf{v}}_{1,9:10}, \text{unif}(\mathcal{A}_v, 2), -\check{\mathbf{v}}_{2,13:14}, \check{\mathbf{v}}_{2,15:16}, \text{rep}(0, 9))^{\mathrm{T}},$$
$$\mathbf{v}_k = \check{\mathbf{v}}_k / \|\check{\mathbf{v}}_k\| \text{ for } k = 1, 2, 3,$$

where $\text{unif}(\mathcal{A}, b)$ denotes a vector of length b whose entries are IID samples from the uniform distribution on the set of real values $\mathcal{A}$, $\mathcal{A}_u = [-1, -0.3] \cup [0.3, 1]$, $\mathcal{A}_v = \pm 1$ and $\check{\mathbf{v}}_{k,a:b}$ denotes the ath–bth entries of $\check{\mathbf{v}}_k$.

(b) Model II ($p = q = 60, n = 50, r^* = 3$): we use the same setting as in model I, except that each 25×1 singular vector in model I is appended with 35 0s to make a 60×1 vector.

The data matrix $\mathbf{Y}$ is then generated by $\mathbf{Y} = \mathbf{XC} + \mathbf{E}$, where the elements of $\mathbf{E}$ are IID samples from $N(0, \sigma^2)$, and σ is chosen to make different SNRs calculated on the basis of the third layer and the noise matrix $\mathbf{E}$. In each setting we replicated the experiment 100 times. The non-zero entries of $\mathbf{v}_k$s have some positional overlap with each other, and the non-zero entries of $\mathbf{u}_k$s take distinct values, some of which can be quite small. These settings make the estimation quite challenging.

For the SEA and IEEA methods, the regularization parameters are chosen on the basis of BIC, so they are both tuned towards sparsity recovery and model selection; for speeding up computation, the IEEA algorithm was carried out with only three iterates. (We have also compared

Table 2. Performances of the IEEA and SEA on sparse SVD recovery

Model		*SNR*	*Results for IEEA*				*Results for SEA*			
			Layer 1	*Layer 2*	*Layer 3*	*Overall*	*Layer 1*	*Layer 2*	*Layer 3*	*Overall*
I	FDR (%)	0.125	1.9	2.7	5.8	3.5	8.3	9.0	9.1	9.9
		0.25	0.9	1.0	2.7	1.6	4.6	6.8	7.9	6.8
		0.5	1.1	1.4	2.2	1.6	7.4	11.7	8.0	10.0
		1	0.9	0.9	1.2	1.0	11.0	17.6	11.2	14.5
	FNR (%)	0.125	0.0	1.6	14.4	5.3	0.0	3.1	15.5	6.2
		0.25	0.0	0.2	2.5	0.9	1.0	3.6	5.0	3.2
		0.5	0.0	0.0	1.3	0.4	0.0	0.5	1.9	0.8
		1	0.0	0.0	0.3	0.1	0.0	1.9	2.4	1.4
II	FDR (%)	0.125	6.4	10.6	11.4	10.4	7.7	15.6	13.4	13.4
		0.25	0.7	5.1	4.9	3.9	17.3	17.0	12.8	17.7
		0.5	2.9	2.0	2.3	2.6	16.8	21.7	17.1	20.5
		1	5.7	4.3	2.1	4.5	18.4	21.7	16.2	20.6
	FNR (%)	0.125	1.1	4.4	10.1	5.2	0.3	3.9	9.9	4.7
		0.25	0.0	0.5	2.9	1.2	0.0	2.5	4.8	2.4
		0.5	0.0	0.1	0.8	0.3	0.0	1.5	2.3	1.2
		1	0.0	0.1	0.3	0.1	0.0	0.1	0.3	0.1

the proposed BIC with fivefold CV for the IEEA estimation; the results are not shown here. In general BIC yields a slightly better selection performance whereas its performance on estimation or prediction is comparable with that of CV, which confirms that the suggested BIC works well.) In contrast, the RRR and NNP methods are tuned for prediction. The RRR estimators of various ranks and NNP estimators over a very fine grid of tuning parameters are first computed. In particular, we use the accelerated proximal gradient algorithm that was proposed by Toh and Yun (2009) for NNP estimation. The rank of the RRR estimator and the tuning parameter of the NNP estimator are then selected on the basis of the best prediction accuracy evaluated on a very large independently generated validation data set, of size $n = 1000$ (Bunea *et al.*, 2011). For each method, the model accuracy is measured by the average SMSE from all runs, i.e. $\mathrm{SMSE} = 100\|\mathbf{C} - \hat{\mathbf{C}}\|_{\mathrm{F}}^2/pq$ for estimation (error Er-Est), and $\mathrm{SMSE} = 100\|\mathbf{XC} - \mathbf{X}\hat{\mathbf{C}}\|_{\mathrm{F}}^2/nq$ for prediction (error Er-Pred).

Table 2 reports the performances on sparse SVD recovery by the SEA and IEEA methods. Since model II involves a much larger number of irrelevant responses or predictors than model I, the recovery of the former is more difficult. Overall the IEEA method performs very well in terms of having low false discovery rates FDR and well-controlled false negative rates FNR. Not surprisingly, the SEA method performs much worse. Its FDRs are much higher than those of the IEEA method because of its inability to distinguish the different SVD layers, and its FDRs do not seem to decrease as the SNR increases for both model I and model II.

We then investigate the prediction and estimation accuracy of the SEA, IEEA, OLS, RRR and NNP estimators. The simulation results are shown in Table 3. It can be seen that, for both model I and model II, the IEEA method performs the best, followed by the SEA, RRR and NNP methods, in this order. The excellent estimation performance of the IEEA and SEA is due to their capability of response or predictor selection, and this property is especially useful when the model dimension is high and the number of irrelevant responses or predictors is large.

Table 3. Estimation and prediction accuracy of the IEEA, SEA, OLS, RRR and NNP estimators

Model	*SNR*	*Error*	*Results for the following methods:*				
			IEEA	*SEA*	*OLS*	*RRR*	*NNP*
I	0.125	Er-Est	3.29	4.28	55.32	9.90	10.99
		Er-Pred	52.10	67.37	416.34	106.40	139.10
	0.25	Er-Est	1.10	1.52	27.21	4.56	6.65
		Er-Pred	17.68	23.20	197.12	46.16	74.55
	0.5	Er-Est	0.57	0.99	15.12	2.44	3.75
		Er-Pred	9.48	15.03	110.45	25.89	43.74
	1	Er-Est	0.23	0.52	7.41	1.17	2.19
		Er-Pred	4.01	7.59	52.95	11.87	23.23
II	0.125	Er-Est	0.52	0.51	51.50	4.87	5.01
		Er-Pred	12.86	14.84	342.59	60.12	72.00
	0.25	Er-Est	0.15	0.29	28.09	3.89	4.22
		Er-Pred	4.47	7.85	176.22	23.64	45.44
	0.5	Er-Est	0.06	0.17	14.81	3.23	3.67
		Er-Pred	1.80	4.27	84.94	9.67	26.87
	1	Er-Est	0.03	0.10	8.40	2.75	3.13
		Er-Pred	0.84	2.35	42.57	4.55	15.43

4.3. Misspecification of the rank

In practice, the rank of $\mathbf{C}$ may be misspecified. For cases when the specified rank, which is denoted as r_0, is larger than the true rank r^*, the penalized sum-of-squares criterion (1.3) may shrink any redundant SVD layer to a zero matrix and hence possibly reduce the rank of $\hat{\mathbf{C}}$ to r^*, owing to its shrinkage effect on the singular values. In contrast, for cases with $r_0 < r^*$, it is expected that the sparsity-inducing criterion will lead to the recovery of the first r_0 dominating layers, i.e. layers with the largest r_0 singular values of the true sparse SVD of $\mathbf{C}$.

We examine the robustness of the proposed method against rank misspecification via simulation, using models I and II in the preceding section, but now the rank is misspecified to be either $r_0 = 2$ or $r_0 = 4$. We performed the IEEA estimation with rank r_0, which is denoted as IEEA(r_0), and used the rank r_0 least squares estimator as the initial estimator. For simplicity we ran the IEEA for six iterations, with the regularization parameters chosen by BIC. For comparison, we also performed the RRR analysis with the rank determined by fivefold CV and the maximum rank set to r_0, which is denoted as RRR(CV, max $= r_0$). The experiment was replicated 100 times for each SNR-level.

Table 4 summarizes the simulation results. We report the frequency of the estimated rank being r^* when $r_0 > r^*$, and the frequency of the estimated rank being r_0 when $r_0 < r^*$ (Rank); the model accuracy is again measured by the average SMSE from all runs as defined earlier (Er-Est and Er-Pred). FDR and FNR for the IEEA method are calculated on the basis of the

Table 4. Performances of the IEEA when the rank is misspecified

Model	*SNR*	*Rank*	*FDR*	*FNR*	*Er-Est*	*Er-Pred*	*Rank*	*Er-Est*	*Er-Pred*
		IEEA(4)					*RRR(CV, max = 4)*		
I	0.125	3 (62%)	5.7%	14.5%	4.3	82.4	3 (53%)	10.4	124.1
	0.25	3 (64%)	7.0%	3.8%	1.8	33.1	3 (100%)	4.8	50.7
	0.5	3 (87%)	2.4%	0.3%	0.5	8.9	3 (100%)	2.3	24.1
	1	3 (90%)	1.4%	0.1%	0.2	4.2	3 (100%)	1.1	12.2
		IEEA(2)					*RRR(CV, max = 2)*		
	0.125	2 (100%)	3.4%	1.4%	6.3	123.2	2 (98%)	11.1	157.9
	0.25	2 (100%)	3.0%	0.4%	5.0	110.4	2 (100%)	7.6	126.6
	0.5	2 (100%)	3.2%	0.4%	4.7	98.8	2 (100%)	6.1	104.7
	1	2 (100%)	4.6%	0.8%	4.6	97.9	2 (100%)	5.3	100.0
		IEEA(4)					*RRR(CV, max = 4)*		
II	0.125	3 (38%)	14.6%	16.2%	0.7	27.3	3 (62%)	4.8	44.8
	0.25	3 (49%)	14.3%	6.7%	0.3	13.0	3 (93%)	3.8	21.5
	0.5	3 (64%)	12.2%	1.3%	0.1	2.9	3 (96%)	3.1	10.2
	1	3 (83%)	5.9%	0.2%	0.0	1.1	3 (97%)	2.8	4.8
		IEEA(2)					*RRR(CV, max = 2)*		
	0.125	2 (100%)	3.2%	1.0%	0.9	45.9	2 (89%)	5.0	74.5
	0.25	2 (100%)	2.0%	0.3%	0.8	38.9	2 (100%)	4.0	46.8
	0.5	2 (100%)	4.3%	0.5%	0.8	39.9	2 (94%)	3.6	47.0
	1	2 (100%)	6.6%	0.1%	0.8	39.8	2 (99%)	3.3	42.0

r_0 estimated SVD layers and their corresponding true counterparts with the largest r_0 singular values (true redundant singular vectors are zero vectors).

It can be seen that our proposed method is robust against rank misspecification. When $r_0 = 4$, the redundant fourth layer is often penalized to be a zero matrix so that the frequencies of recovering the true rank are generally high. When $r_0 = 2$, our method always leads to the recovery of the first two dominating layers of the true sparse SVD. For cases with low SNR, our method may have better rank recovery or discovery performance than the RRR method; for cases with moderate to high SNR, the RRR method shows better rank determination performance especially for $r_0 = 4$. However, our method always outperforms the RRR method in both estimation and prediction, and further examination shows that, even if our method may occasionally fail to eliminate the redundant layer, the estimated redundant layer is generally very sparse and weak, rendering our method immune to excessive overfitting. We also point out that the estimation, prediction and variable selection performances of the method proposed are, as expected, all slightly worse than the case of known rank; see Tables 2 and 3.

4.4. Biclustering: lung cancer data

We illustrate by a real application the effectiveness of the proposed method in microarray biclustering analysis (Busygin *et al.*, 2008). The goal is to identify sets of biologically relevant genes that are expressed at different levels in different types of cancer by using microarray gene expression data, in which usually thousands of genes are measured for only a few subjects. The method proposed is well suited for such a simultaneous selection problem. We show that a special case of the proposed regression method with an identity design matrix can serve as an efficient biclustering tool. Our method is flexible and more general in that it also allows easy incorporation of the cancer group information, if available, to perform a supervised search of the gene clusters. Moreover, the method can be further extended to adjust for 'unwanted' expression heterogeneity, on which a general statistical microarray biclustering method can be built.

The gene expression data that we consider here consist of expression levels of $q = 12625$ genes, measured from $n = 56$ subjects. Among the 56 subjects, 17 of them were known to be normal (Normal), and the remaining 39 were known to be with one of three types of lung cancer: 20 of them were with pulmonary carcinoid tumours (Carcinoid), 13 with colon metastases (Colon) and six with small cell carcinoma (SmallCell). The data form an $n \times q$ matrix $\mathbf{Y}$ whose rows represent the subjects, grouped sequentially by the type of cancer (Carcinoid, Colon, Normal and SmallCell), and the columns correspond to the genes. A more detailed description of the data can be found in Bhattacharjee *et al.* (2001). The data were analysed by Liu *et al.* (2008) and more recently by Lee *et al.* (2010), in which the SSVD method was proposed for biclustering.

On letting the covariate matrix $\mathbf{X}$ be the $n \times n$ identity matrix, our sparse reduced rank regression model reduces to a low rank matrix approximation problem for $\mathbf{Y}$, which can serve as an unsupervised learning tool for biclustering since the available cancer type information is not used. In this special case, our method shares a similar idea with the SSVD method in Lee *et al.* (2010). Not surprisingly, our estimation result is also similar to that of the SSVD method. Hence we omit the detailed estimation results, although it is worth noting that our proposed method is indeed capable of simultaneously linking sets of genes to sets of subjects, and the associations between gene groups and types of cancer are clearly revealed in three identified sparse SVD layers. Heat maps of the original gene expression matrix, the reduced rank estimate and the three estimated layers are plotted in Fig. 1. To visualize the gene clustering better,

(a) all entries of the plotted matrices are divided by the maximum absolute value of the entries,

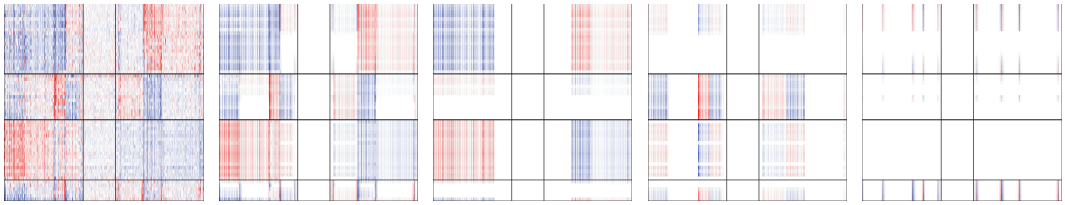

Fig. 1. Heat maps of (a) the original gene expression matrix, (b) the reduced rank estimate and (c), (d), (e) the three estimated SVD layers by unsupervised biclustering: the 5200 selected genes and the other 1000 randomly chosen unselected genes are plotted; all entries of the matrices plotted are divided by the maximum absolute value of the entries; the genes are sorted hierarchically (firstly the genes are sorted on the basis of the ascending order of the entries of $\hat{\mathbf{v}}_1$, which automatically forms three gene groups according to the sign of the entries; secondly, within each group, the genes are sorted on the basis on $\hat{\mathbf{v}}_2$, and then nine gene groups are formed; finally, the sorting procedure is repeated on the basis of $\hat{\mathbf{v}}_3$; the horizontal lines in each panel reveal the four types of cancer of the subjects (Carcinoid, Colon, Normal and SmallCell from top to bottom); the vertical lines in each panel reveal the 1000 unselected genes at the second column

(b) only the 5200 selected genes plus 1000 randomly chosen unselected genes are plotted and
(c) the genes in Fig. 1 are sorted hierarchically.

The 1000 unselected genes are included to show that the zero-out areas in the estimated SVD layers indeed correspond to non-informative areas of the original gene expression matrix. A very strong contrast between the Carcinoid group and the Normal group can be seen from the first estimated SVD layer, and another strong contrast between the Colon group and the Normal group can be seen from the second layer. However, because of a failure in accounting for within-group variations, the unsupervised learning may also provide irrelevant or even inconsistent information about gene–cancer associations, as also found in Lee *et al.* (2010). In particular, in the third estimated SVD layer, some of the subjects of the Carcinoid group had positive responses, whereas others in this group did not respond or even showed opposite responses. Although such information may be valuable in that it suggests possible subgroup structure in the Carcinoid group, it is irrelevant on how to distinguish the four known categories.

If the cancer type information is available as for this data set, such information can be incorporated for supervised learning of the gene clusters and their contrasts across different types of cancer. This can be done by constraining the left singular vectors to be linear combinations of the dummy variables of the types of cancer, i.e. we fit the data by model (1.1) with a 56×4 covariate matrix

$$\mathbf{X} = \begin{pmatrix} (1/\sqrt{20})\mathbf{1}_{20} & 0 & 0 & 0 \\ 0 & (1/\sqrt{13})\mathbf{1}_{13} & 0 & 0 \\ 0 & 0 & (1/\sqrt{17})\mathbf{1}_{17} & 0 \\ 0 & 0 & 0 & (1/\sqrt{6})\mathbf{1}_{6} \end{pmatrix},$$

where, for instance, $\mathbf{1}_{20}$ is a 20×1 vector consisting of 1s. The coefficient matrix $\mathbf{C}$ is a $4 \times 12\,625$ matrix, and each of its 1×4 left singular vectors can be interpreted as group effects rather than individual subject effects, whereas each of its right singular vectors still represents the gene effects. Since $\mathbf{X}$ is still orthogonal, computation remains fast. This model implies that, for the rank 3 SVD approximation of $\mathbf{Y}$, the left singular vectors equal the products of $\mathbf{X}$ times the left singular vectors of $\mathbf{C}$; hence subjects of identical type of cancer enjoy the same mean structure. Extending the model to allow mixed effects in the gene–cancer associations is an interesting future research problem.

We then perform the supervised learning by using the IEEA method with BIC. We consider

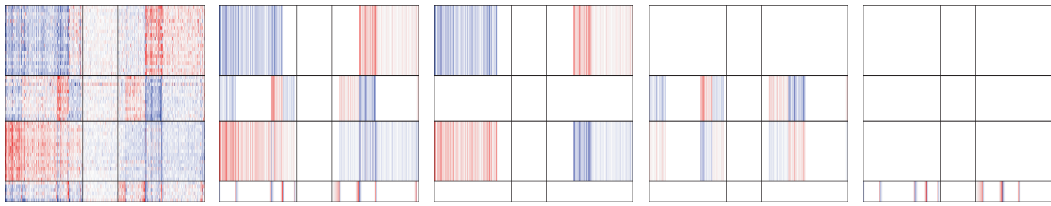

Fig. 2. Heat maps of (a) the original gene expression matrix, (b) the reduced rank estimate and (c), (d), (e) the three estimated SVD layers by supervised learning: all set-ups are the same as in Fig. 1

the first three layers since $\mathbf{Y}$ is centred so that it is orthogonal to $\mathbf{1}_{56}$, but $\mathbf{1}_{56}$ lies in the column space of $\mathbf{X}$; in fact, the four estimated singular values are 200.22, 119.27, 77.83 and 0.000065. Only 4663 genes are selected overall. Among those selected, 3507, 2231 and 1089 genes are involved in the three layers. Heat maps of the original gene expression matrix, the reduced rank estimate and the three estimated layers are plotted in Fig. 2. By supervised learning, more than 1000 genes are further eliminated in the three layers compared with unsupervised learning, and only information about gene–cancer type associations are extracted and kept. Yet our sparse estimator still reveals the primary chequered structure in the original gene expression matrix as shown in Figs 2(a) and 2(b). The first SVD layer presents a strong contrast between the Carcinoid group and the Normal group, the second layer clearly presents a contrast between the Colon group and the Normal group and the third layer singles out the SmallCell group. Some weaker contrasts that were previously seen in the unsupervised learning have been eliminated since they might be explained by within-group variations. In particular, in the third SVD layer, the subgroup structure that was found earlier in the Carcinoid group is now absent from the supervised learning results because it is irrelevant with respect to the known group information.

In gene expression studies, expression heterogeneity due to technical, genetic, environmental or demographic variables is very common (Leek and Storey, 2007). It is desirable to adjust for these covariate effects or 'unwanted' variations while studying the clustering with respect to the primary variable, e.g. type of cancer. This can be done via a reduced rank regression model with two sets of regressors: $\mathbf{Y} = \mathbf{XC} + \mathbf{ZG} + \mathbf{E}$ where $\mathbf{X}$ is an $n \times p$ matrix constructed from the primary variable such as cancer group information, $\mathbf{C}$ is a $p \times q$ matrix which may admit a reduced rank sparse SVD structure, $\mathbf{Z}$ is an $n \times l$ matrix consisting of additional l (confounding) variables measured on the subjects, $\mathbf{G}$ is an $l \times q$ coefficient matrix that may be of full rank and the other terms are defined as in model (1.1). This type of model formation was first suggested in the seminal work of Anderson (1951) and was studied by Reinsel and Velu (1998), chapter 3, under the classical least squares setting. Here, under our regularized regression framework, the above extension adds no significant difficulty in estimation. One could still use a block coordinate descent algorithm to update $\mathbf{C}$ and $\mathbf{G}$ iteratively until convergence. We shall report investigations of this approach elsewhere.

5. Theoretical properties

We state the main theoretical results and outline their proofs, leaving details and further results to the on-line supplementary materials. In this section, the regularization parameters are assumed to be a function of the sample size, and hence are written as $\lambda_k^{(n)}$. The following conditions are needed in the theoretical developments.

Condition 1. $(1/n)\mathbf{X}^{\mathrm{T}}\mathbf{X} \to \Gamma$ almost surely as $n \to \infty$, where Γ is a fixed, positive definite matrix.

Condition 2. The errors $\mathbf{e}_i$ $(i=1,\ldots,n)$ are IID with $E(\mathbf{e}_i)=\mathbf{0}$ and $\mathrm{cov}(\mathbf{e}_i)=\Sigma$, a positive definite matrix.

Condition 3. $\lambda_k^{(n)}/\sqrt{n}\to 0$ and $(\lambda_k^{(n)}/\sqrt{n})n^{\gamma/2}\to\infty$ as $n\to\infty$, for $k=1,\ldots,r^*$, with $\gamma>0$ being prespecified.

In what follows, let $\mathbf{Z}_{\mathcal{AB}}$ denote a submatrix of an arbitrary matrix $\mathbf{Z}$ whose rows and columns are chosen from $\mathbf{Z}$ according to some index sets $\mathcal{A}$ and $\mathcal{B}$ respectively. For simplicity, we may write $\mathbf{Z}_{\mathcal{AB}}=\mathbf{Z}_{\cdot\mathcal{B}}$ and $\mathbf{Z}_{\mathcal{AB}}=\mathbf{Z}_{\mathcal{A}\cdot}$ when respectively $\mathcal{A}$ and $\mathcal{B}$ consist of all the row and column indices.

The set of all $p\times q$ matrices of rank smaller than or equal to r $(r\leqslant p,q)$, which is denoted as $\Omega^{(r)}$, admits a manifold structure. Any matrix $\mathbf{Z}\in\Omega^{(r)}$ can be written as a product $\mathbf{Z}=\mathbf{U}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}$ is a $p\times r$ matrix and $\mathbf{V}$ is a $q\times r$ rank r matrix. This decomposition is not unique since $\mathbf{Z}=\mathbf{U}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{V}^{\mathrm{T}}$ for any $r\times r$ invertible matrix $\mathbf{Q}$. However, since $\mathrm{rank}(\mathbf{V})=r$, there is an $r\times r$ submatrix $\mathbf{V}_{\mathcal{L}\cdot}$ of $\mathbf{V}$, whose rows are linearly independent and hence is invertible. It then follows that $\mathbf{V}\mathbf{V}_{\mathcal{L}\cdot}^{-1}$ has a submatrix that is the r-dimensional identity matrix $\mathbf{I}_r$. On the basis of this observation, $\Omega^{(r)}$ can be presented as a manifold that is a union of $\binom{q}{r}$ subsets or charts each of which admits a Euclidean co-ordinate system, i.e. $\Omega^{(r)}=\cup_{\mathcal{L}\in\Pi}\Omega_{\mathcal{L}}^{(r)}$, where Π consists of all size r subsets of the set $\{1,\ldots,q\}$ and

$$\Omega_{\mathcal{L}}^{(r)}=\{\mathbf{U}\mathbf{V}^{\mathrm{T}};\mathbf{U}\text{ is a }p\times r\text{ matrix and }\mathbf{V}\text{ is a }q\times r\text{ matrix with }\mathbf{V}_{\mathcal{L}\cdot}=\mathbf{I}_r\}.$$

Now suppose that the true model is given by expression (1.1), where the rank of the coefficient matrix $\mathbf{C}$ has been correctly identified to be r^*. Let $\mathbf{C}=\mathbf{U}^*\mathbf{V}^{*\mathrm{T}}$, where $\mathbf{U}^*=(u_{ik}^*)_{p\times r^*}$ is a $p\times r^*$ orthogonal matrix and $\mathbf{V}^*=(v_{jk}^*)_{q\times r^*}$ is a $q\times r^*$ orthogonal matrix. Here for simplicity the singular values have been absorbed into the singular vectors. Because the parameter space $\Omega^{(r^*)}$ is a manifold, without loss of generality, we can assume the true coefficient matrix $\mathbf{C}\in\Omega_{\mathcal{L}}^{(r^*)}$ where $\mathcal{L}=\{l_1,\ldots,l_{r^*}\}$ is a fixed size r^* index set.

Let $Q_n(\mathbf{U},\mathbf{V})$ denote the objective function as in equation (1.3), i.e.

$$Q_n(\mathbf{U},\mathbf{V})=\tfrac{1}{2}\|\mathbf{Y}-\mathbf{X}\mathbf{U}\mathbf{V}^{\mathrm{T}}\|_{\mathrm{F}}^2+\sum_{k=1}^{r^*}\left(\lambda_k^{(n)}\sum_{i=1}^{p}\sum_{j=1}^{q}w_{ijk}|u_{ik}v_{jk}|\right)\tag{5.1}$$

and let $(\hat{\mathbf{U}}^{(n)},\hat{\mathbf{V}}^{(n)})=\arg\min\{Q_n(\mathbf{U},\mathbf{V})\}$.

Theorem 1 (existence of a local minimum). Suppose that conditions 1 and 2 are satisfied, and suppose that $\lambda_k^{(n)}/\sqrt{n}\to\lambda_k\geqslant 0$ as $n\to\infty$ for $k=1,\ldots,r^*$. Then there is a local minimizer $(\hat{\mathbf{U}}^{(n)},\hat{\mathbf{V}}^{(n)})$ of $Q_n(\mathbf{U},\mathbf{V})$, such that $\|\hat{\mathbf{U}}^{(n)}-\mathbf{U}^*\|=O_p(n^{-1/2})$ and $\|\hat{\mathbf{V}}^{(n)}-\mathbf{V}^*\|=O_p(n^{-1/2})$.

Theorem 2 (asymptotic normality). Suppose that conditions 1–3 are satisfied. Let vector $\hat{\mathbf{u}}_{\mathcal{A}}^{(n)}$ and $\mathbf{u}_{\mathcal{A}}^*$ collect all entries in respectively $\hat{\mathbf{U}}^{(n)}$ and $\mathbf{U}^*$ corresponding to the non-zero elements in $\mathbf{U}^*$, and let vector $\hat{\mathbf{v}}_{\mathcal{B}}^{(n)}$ and $\mathbf{v}_{\mathcal{B}}^*$ respectively collect all entries in $\hat{\mathbf{V}}^{(n)}$ and $\mathbf{V}^*$ corresponding to the non-zero elements in $\mathbf{V}^*$. Then

(a) $(\hat{\mathbf{u}}_{\mathcal{A}}^{(n)}-\mathbf{u}_{\mathcal{A}}^*)\sqrt{n}$ and $(\hat{\mathbf{v}}_{\mathcal{B}}^{(n)}-\mathbf{v}_{\mathcal{B}}^*)\sqrt{n}$ are both asymptotically normally distributed with zero mean and

(b) $(\hat{\mathbf{u}}_{\mathcal{A}^{\mathrm{c}}}^{(n)}-\mathbf{u}_{\mathcal{A}^{\mathrm{c}}}^*)\sqrt{n}\to_{\mathrm{d}}0$ and $(\hat{\mathbf{v}}_{\mathcal{B}^{\mathrm{c}}}^{(n)}-\mathbf{v}_{\mathcal{B}^{\mathrm{c}}}^*)\sqrt{n}\to_{\mathrm{d}}0$ as $n\to\infty$.

Theorem 3 (selection consistency). Suppose that conditions 1–3 are satisfied. Let $\mathcal{A}=\{(i,k):u_{ik}^*\neq 0\}$ and $\mathcal{B}=\{(j,k):v_{jk}^*\neq 0\}$, and let $\mathcal{A}^{(n)}=\{(i,k):\hat{u}_{ik}^{(n)}\neq 0\}$ and $\mathcal{B}^{(n)}=\{(j,k):\hat{v}_{jk}^{(n)}\neq 0\}$. Then $P(\mathcal{A}^{(n)}=\mathcal{A})\to 1$ and $P(\mathcal{B}^{(n)}=\mathcal{B})\to 1$ as $n\to\infty$.

We outline the key steps in proving the preceding theorems.

(a) We start from constructing the adaptive weights w_{ijk} based on the SVD of some initial estimator of $\mathbf{C}$, as shown in equation (2.2). In particular, the rank r^* least squares estimator of $\mathbf{C}$ is explicitly given by $\tilde{\mathbf{C}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{H}\mathbf{H}^{\mathrm{T}}$ where $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{r^*})$ and $\mathbf{h}_k$ is the normalized eigenvector that corresponds to the kth largest eigenvalue of the matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. It is shown that $\mathrm{vec}(\tilde{\mathbf{C}} - \mathbf{C})\sqrt{n} \to_{\mathrm{d}} N(\mathbf{0}, \Sigma_{\mathrm{c}})$ where Σ_{c} is explicitly given by expression (2.36) in Reinsel and Velu (1998). Recall that $\mathbf{C} = \Sigma_{k=1}^{r^*} d_k \mathbf{u}_k \mathbf{v}_k^{\mathrm{T}}$ is the SVD of $\mathbf{C}$, where $d_1 > \ldots > d_{r^*} > 0$. Similarly, let $\tilde{d}_k$, $\tilde{\mathbf{u}}_k$ and $\tilde{\mathbf{v}}_k$ $(k = 1, \ldots, r^*)$ be the singular values and left and right singular vectors respectively of $\tilde{\mathbf{C}}$. Then using the perturbation expansion of matrices (theorem 3, Izenman (1975)), we showed that $(\tilde{d}_k - d_k)\sqrt{n}$, $(\tilde{\mathbf{u}}_k - \mathbf{u}_k)\sqrt{n}$ and $(\tilde{\mathbf{v}}_k - \mathbf{v}_k)\sqrt{n}$ for $k = 1, \ldots, r^*$ are jointly asymptotically normally distributed with zero mean. It turns out that such adaptive weights as designed in equation (2.2) play an important role in achieving selection consistency.

(b) We have assumed that $\mathbf{C} \in \Omega_{\mathcal{L}}^{(r^*)}$ for a fixed size r^* index set $\mathcal{L}$. This is equivalent to setting the submatrix $\mathbf{V}^*_{\mathcal{L}\cdot}$ of $\mathbf{V}^*$, which is also denoted as $\mathbf{Q}$, to be invertible. Denote $\check{\mathbf{U}} = \mathbf{U}^*\mathbf{Q}^{\mathrm{T}}$ and $\check{\mathbf{V}} = \mathbf{V}^*\mathbf{Q}^{-1}$, so that $\check{\mathbf{V}}_{\mathcal{L}\cdot} = \mathbf{I}_{r^*}$ and $\mathbf{C} = \mathbf{U}^*\mathbf{V}^{*\mathrm{T}} = \check{\mathbf{U}}\check{\mathbf{V}}^{\mathrm{T}}$. A neighbourhood $\mathcal{N}(\mathbf{C}, h_n)$, which is of radius $h_n = O(\sqrt{[\log\{\log(n)\}]}$ and centred at $\mathbf{C}$, can then be constructed in the chart $\Omega_{\mathcal{L}}^{(r^*)}$, i.e.

$$\begin{aligned}\mathcal{N}(\mathbf{C}, h) = &\left\{\left(\check{\mathbf{U}} + \frac{1}{\sqrt{n}}\check{\mathbf{A}}\right)\left(\check{\mathbf{V}} + \frac{1}{\sqrt{n}}\check{\mathbf{B}}\right)^{\mathrm{T}}; \check{\mathbf{A}} = (\check{a}_{ik})_{p\times r^*} \text{ with } \|\check{\mathbf{A}}\| \leqslant h_n, \check{\mathbf{B}} = (\check{b}_{jk})_{q\times r^*}\right.\\ &\left.\text{with } \check{\mathbf{B}}_{\mathcal{L}\cdot} = \mathbf{0} \text{ and } \|\check{\mathbf{B}}\| \leqslant h_n.\right\}\\ = &\left\{\left(\mathbf{U}^* + \frac{1}{\sqrt{n}}\mathbf{A}\right)\left(\mathbf{V}^* + \frac{1}{\sqrt{n}}\mathbf{B}\right)^{\mathrm{T}}; \check{\mathbf{A}} = (\check{a}_{ik})_{p\times r^*} \text{ with } \|\check{\mathbf{A}}\| \leqslant h_n, \check{\mathbf{B}} = (\check{b}_{jk})_{q\times r^*}\right.\\ &\left.\text{with } \check{\mathbf{B}}_{\mathcal{L}\cdot} = \mathbf{0} \text{ and } \|\check{\mathbf{B}}\| \leqslant h_n.\right\},\end{aligned}$$

where $\mathbf{A} = \check{\mathbf{A}}\mathbf{Q}^{-\mathrm{T}}$ and $\mathbf{B} = \check{\mathbf{B}}\mathbf{Q}$. We then prove theorem, 1 by showing the existence of a local minimum in the interior of $\mathcal{N}(\mathbf{C}, h_n)$, i.e., for any given $\varepsilon > 0$, there is a sufficiently large constant h_n such that

$$P\left[\inf_{\|\check{\mathbf{A}}\| = \|\check{\mathbf{B}}\| = h_n}\left\{Q_n\left(\mathbf{U}^* + \frac{1}{\sqrt{n}}\mathbf{A}, \mathbf{V}^* + \frac{1}{\sqrt{n}}\mathbf{B}\right) > Q_n(\mathbf{U}^*, \mathbf{V}^*)\right\}\right] \geqslant 1 - \varepsilon.$$

(c) To prove theorem 2, we first show that $\Psi_n(\mathbf{A}, \mathbf{B}) \equiv Q_n\{\mathbf{U}^* + (1/\sqrt{n})\mathbf{A}, \mathbf{V}^* + (1/\sqrt{n})\mathbf{B}\} - Q_n(\mathbf{U}^*, \mathbf{V}^*)$ defined in $\mathcal{N}(\mathbf{C}, h_n)$ converges in distribution to a biconvex limit function $\Psi(\mathbf{A}, \mathbf{B})$. Studying the behaviour of $\Psi(\mathbf{A}, \mathbf{B})$ is challenging because of its non-convexity. However, by adopting a profile likelihood approach and through extensive use of matrix algebra, it can be shown that $\Psi(\mathbf{A}, \mathbf{B})$ has a unique minimum. We then prove theorem 2 by applying the 'Argmax' theorem (page 81, van der Vaart (2000)) for $-\Psi_n$.

(d) The objective function (5.1) admits a conditional lasso structure, as similarly shown in Section 2. Theorem 3 is then proved by using the consistency results that were established in theorem 2 and a set of conditional Karush–Kuhn–Tucker optimality conditions derived from the conditional lasso models.

By theorems 2 and 3, we have shown that the proposed fully iterative IEEA method, which aims at solving the general objective function (1.3), achieves the oracle property. In fact, it can also be shown that the EEA method proposed, which is non-iterative and relies on some initial $\sqrt{n}$-consistent estimator to reduce the general problem to multiple parallel unit rank problems,

also enjoys desirable large sample properties, e.g. asymptotic normality and selection consistency. The proofs are very similar to those of the general theory in the unit rank case, except for only a few key differences; for example the Ψ_n-function in the EEA case involves some extra additive terms that are functions of the initial estimator. It turns out that the boundedness of these terms follows easily from the $\sqrt{n}$-consistency of the initial estimator, the perturbation expansion of matrices (theorem 3, Izenman (1975)) and the delta method. We omit the details here.

6. Discussion

There are several potential directions for future research. Firstly, we have mainly considered the adaptive lasso penalty (Zou, 2006). It would be worthwhile to explore other sparsity penalty forms such as the smoothly clipped absolute deviation penalty (Fan and Li, 2001) and the elastic net penalty (Zou and Hastie, 2005). Secondly, the current methodology requires correct determination of the rank of the coefficient matrix. Motivated by Yuan *et al.* (2007), it is interesting to extend our methodology to conduct simultaneous rank determination and sparse coefficient estimation in multivariate modelling. The multiplicative penalty form that we have considered not only promotes sparsity in the singular vectors; it also shrinks the singular values of each SVD layer towards 0. Therefore, a promising approach is to set the rank to an upper bound r ($r^* \leqslant r \leqslant \min(p, q)$) and then to fit the model via the penalized approach developed. Although the singular vectors of the redundant layers are no longer identifiable, the penalty term (1.4) for each $k > r^*$ can be expected to penalize the redundant layer to be exactly a zero matrix, owing to its shrinkage effect on the singular value. Another interesting and pressing problem concerns further extending the methodology and theory to high dimensional situations with $p = p_n \to \infty$ or $q = q_n \to \infty$, as problems involving a huge number of responses or predictors such as microarray analysis and genomewide association studies become increasingly common; see the relevant works in the framework of the lasso and related estimators by Zhao and Yu (2006), Zhang and Huang (2008) and the references therein.

Acknowledgements

The authors thank two referees and the Associate Editor for very helpful comments and are grateful to the US National Science Foundation (NSF-0934617) for partial financial support.

References

Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.*, **22**, 327–351.
Anderson, T. W. (2002) Specification and misspecification in reduced rank regression. *Sankhya* A, **64**, 193–205.
Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*, 3nd edn. New York: Wiley.
Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J. and Meyerson, M. (2001) Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natn. Acad. Sci. USA*, **98**, 13790–13795.
Breheny, P. and Huang, J. (2009) Penalized methods for bi-level variable selection. *Statist. Interface*, **2**, 369–380.
Bunea, F., She, Y. and Wegkamp, M. (2011) Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, **39**, 1282–1309.
Busygin, S., Prokopyev, O. and Pardalos, P. M. (2008) Biclustering in data mining. *Comput. Oper. Res.*, **35**, 2964–2987.
Camba-Mendez, G., Kapetanios, G., Smith, R. J. and Weale, M. R. (2003) Tests of rank in reduced rank regression models. *J. Bus. Econ. Statist.*, **21**, 145–155.
Donoho, D. L. and Johnstone, I. M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Statist. Ass.*, **90**, 1200–1224.
Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.
Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007) Pathwise coordinate optimization. *Ann. Appl. Statist.*, **2**, 302–332.
Friedman, J. H., Hastie, T. and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softwr.*, **33**, 1–22.
Gorski, J., Pfeuffer, F. and Klamroth, K. (2007) Biconvex sets and optimization with biconvex functions: a survey and extensions. *Math. Meth. Oper. Res.*, **66**, 373–407.
Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *J. Multiv. Anal.*, **5**, 248–264.
Lee, M., Shen, H., Huang, J. Z. and Marron, J. S. (2010) Biclustering via sparse singular value decomposition. *Biometrics*, **66**, 1087–1095.
Leek, J. T. and Storey, J. D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet.*, **3**, article e161.
Liu, Y., Hayes, D. N. N., Nobel, A. and Marron, J. S. (2008) Statistical significance of clustering for high-dimension, low sample size data. *J. Am. Statist. Ass.*, **103**, 1281–1293.
Obozinski, G., Wainwright, M. J. and Jordan, M. I. (2008) Union support recovery in high-dimensional multivariate regression. *Technical Report*. Department of Statistics, University of California, Berkeley. (Available from `http://www.stat.berkeley.edu/tech-reports/761.pdf`.)
Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R. and Wang, P. (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Statist.*, **4**, 53–77.
R Development Core Team (2008) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Reinsel, G. C. and Velu, P. (1998) *Multivariate Reduced-rank Regression: Theory and Applications*. New York: Springer.
Schott, J. R. (2005) *Matrix Analysis for Statistics*, 2nd edn. Hoboken: Wiley.
Stenseth, N. C., Jordel, P. E., Chan, K.-S., Hansen, E., Knutsen, H., Andre, C., Skogen, M. D. and Lekve, K. (2006) Ecological and genetic impact of Atlantic cod larval drift in the Skagerrak. *Proc. R. Soc. Lond.* B, **273**, 1085–1092.
Stone, M. (1974) Cross-validation and multinomial prediction. *Biometrika*, **61**, 509–515.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Toh, K.-C. and Yun, S. (2009) An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacif. J. Optim.*, **6**, 615–640.
Turlach, B. A., Venables, W. N. and Wright, S. J. (2005) Simultaneous variable selection. *Technometrics*, **47**, 349–363.
van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge: Cambridge University Press.
Vounou, M., Nichols, T. E. and Montana, G. (2010) Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *NeuroImage*, **53**, 1147–1159.
Witten, D. M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.
Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc.* B, **69**, 329–346.
Zhang, C.-H. and Huang, J. (2008) The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.
Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.
Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.
Zou, H., Hastie, T. and Tibshirani, R. (2007) On the degree of freedom of the lasso. *Ann. Statist.*, **35**, 2173–2192.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials for "Reduced-rank stochastic regression with a sparse singular value decomposition" '.

Please note: Wiley–Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the author for correspondence for the article.