

Reduced-rank Stochastic Regressions with a Sparse Singular-Value-Decomposition

Kun Chen, Kung-sik Chan, Nils Chr. Stenseth

May 15, 2010

Abstract

For a reduced-rank multivariate stochastic regression model of, say, 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 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 the afore-mentioned problem. Computation algorithms and regularization parameter selection methods are developed, and the properties of the new method are explored both theoretically and by simulation. We apply the proposed approach to analyzing the Norwegian Skagerrak coastal cod abundance data for simultaneously capturing the spawning peak and identifying significant North Sea larval drift effects among coastal fjords. We also apply the proposed model to the biclustering problem using microarray data.

1 Introduction

We consider the reduced-rank regression model,

$$\mathbf{s}_t = \mathbf{C}\mathbf{g}_t + \mathbf{e}_t, \quad t = 1, \dots, T, \quad (1.1)$$

where $\mathbf{s}_t = (s_{1t}, \dots, s_{mt})^T$ is an $m \times 1$ vector of response variables, $\mathbf{g}_t = (g_{1t}, \dots, g_{nt}^T)$ is an $n \times 1$ vector of predictor variables, \mathbf{C} is an $m \times n$ regression coefficient matrix with $\text{rank}(\mathbf{C}) = r \leq \min(m, n)$, and $\mathbf{e}_t = (e_{1t}, \dots, e_{mt})^T$ is the $m \times 1$ vector of random errors, which is assumed to be independently and identically distributed (*i.i.d.*) with mean vector $E(\mathbf{e}_t) = 0$ and covariance matrix $\text{Cov}(\mathbf{e}_t) = \Sigma_e$, an $m \times m$ positive-definite matrix. We assume T observations are available, and define the $m \times T$ data matrix $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_T]$, the $n \times T$ covariate matrix $\mathbf{G} = [\mathbf{g}_1, \dots, \mathbf{g}_T]$ and the $m \times T$ error matrix $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_T]$. The model in terms of the complete data can be written as

$$\mathbf{S} = \mathbf{C}\mathbf{G} + \mathbf{E}.$$

The unknown parameters in the above model are the rank r regression coefficient matrix \mathbf{C} and the error covariance matrix Σ_e . The classical reduced-rank regression ~~criterion~~ is given by

$$tr[\Gamma(\mathbf{S} - \mathbf{CG})(\mathbf{S} - \mathbf{CG})^T] \quad (1.2)$$

where \mathbf{C} is restricted to be a ~~reduced-rank~~ matrix, and Γ is an $m \times m$ positive definite matrix usually ~~choosing~~ to be ~~an~~ identity matrix or $\tilde{\Sigma}_e^{-1}$, where $\tilde{\Sigma}_e$ is some initial estimates of Σ_e . In practice, a ~~legitimate~~ estimate of the covariance matrix that is positively definite might be hard to obtain, especially for high dimensional data. Here we mainly focus on the case when Γ is identity matrix. The methodology can be easily extended to the ~~latter~~ case.

~~One immediate~~ problem of the reduced-rank regression is the identification of the rank of the regression coefficient matrix. It is well-known that principal component analysis and canonical correlation analysis can be regarded as special cases of reduced-rank regression (Izenman, 1975). The relationship between canonical correlation analysis and reduced-rank regression enables us to estimate the rank by testing whether certain correlations are zero, which leads to the likelihood ratio test for testing the hypothesis that $rank(\mathbf{C}) = r$, see Anderson and Anderson (1984). Hence, an approach to identify the rank is to adopt the smallest value of r for which $H_0 : rank(\mathbf{C}) = r$ is not rejected. Other tools for the specification of the rank include the AIC criterion (Akaike, 1974), the BIC criterion (Schwarz, 1978) and cross-validation (Stone, 1974), based on the predictive performance of models of various ranks. More recently, Yuan et al. (2007) proposed a novel penalized least squares approach to conduct dimension reduction and coefficient estimation simultaneously in the multivariate linear model. The penalty they considered encourages ~~the sparsity among~~ singular values so that the rank can ~~be~~ automatically ~~be~~ determined as the number of nonzero singular values. In this paper, we assume the rank of the coefficient matrix has been correctly identified, and our goal is to ~~improve~~ the coefficient matrix ~~estimation~~.

The rank- r regression coefficient matrix \mathbf{C} 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, i.e.

$$\mathbf{C} = \mathbf{UDV}^T = \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^T = \sum_{k=1}^r \mathbf{C}_k \quad (1.3)$$

where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_r]$ consists of r left singular vectors, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_r]$ consists of r right singular vectors, $\mathbf{D} = diag(d_1, \dots, d_r)$ is a diagonal matrix with positive singular values $d_1 \geq \dots \geq d_r$ on its diagonal, and $\mathbf{C}_k = d_k \mathbf{u}_k \mathbf{v}_k^T$ is the layer- k unit-rank matrix of \mathbf{C} . This singular value decomposition (SVD) representation shows that \mathbf{C} is composed of r orthogonal layers of ~~different~~ importance, and each layer provides a distinct channel relating the response variables to the explanatory variables.

For facilitating interpretation, it is often desirable that these left and right singular vectors be sparse or enjoy some smoothness property. This is motivated by two applications ~~we consider in the article~~. The first is an ecological application, in which we analyze the Norwegian Skagerrak coastal cod abundance data for simultaneously

capturing the spawning peak and identifying significant North Sea larval drift effects among coastal fjords. It is hypothesized that among 18 coastal fjords under consideration, only the fjords which are exposed to the North Sea could potentially receive larva drift from ~~outside sources~~. Hence the left singular vector, which turns out to represent the larval drift effects, is believed to be sparse. ~~In the mean time~~, the right singular vector, which represent the spawning ~~effects~~ over a 45-day period, is believed to be smooth in time and peak at some point in between. In the second application, the goal is to identify sets of biologically relevant genes that are significantly expressed for certain cancer types using microarray expression data. The data consist of expression levels of thousands of genes, measured from a much smaller number of subjects, who are known to be either normal subjects or patients with different types of cancer. ~~To~~ be able to simultaneously identify related genes and subject groups, ~~making~~ use of the grouping information, adjusting for the covariate effects, and promoting sparsity in estimation ~~are all very important~~.

We propose a regularized reduced-rank regression approach for solving the aforementioned problems. To induce sparsity in a singular vector, a suitable penalty term, e.g. a multiple of its L_1 norm, could be added to the minimization objective in (1.2). In some applications, there are cases when the right singular vectors are believed to be smooth in some known covariate \mathbf{h} . Under such circumstance, a suitable smoothing basis can be used to expand the right singular matrix \mathbf{V} , i.e. $\mathbf{V} = \mathbf{Q}\mathbf{V}^*$, where \mathbf{Q} is an $n \times n^*$ transformation matrix whose columns consist of basis functions evaluated at each h_t , and \mathbf{V}^* is the $n^* \times r$ transformed coefficient matrix on which sparsity penalty can then be imposed. By redefining \mathbf{G} to be $\mathbf{Q}^T\mathbf{G}$, the model reduces to the case of sparsity. Without lost of generality, here 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 $k = 1, \dots, r$:

$$\frac{1}{2} \text{tr}\{[\mathbf{S} - (\sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^T) \mathbf{G}] [\mathbf{S} - (\sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^T) \mathbf{G}]^T\} + \sum_{k=1}^r Pe(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)), \quad (1.4)$$

where $Pe(\cdot)$ is some penalty function, and λ_k s are the regularization parameters controlling the degrees of penalization.

To prompt sparsity in \mathbf{u}_k and \mathbf{v}_k , we consider the class of adaptive lasso penalties (Zou, 2006). Specifically, we consider

$$Pe(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)) = \lambda_k \sum_{i=1}^m \sum_{j=1}^n w_{ijk} |d_k u_{ik} v_{jk}|, \quad (1.5)$$

where the w_{ijk} s are possibly data-driven weights, which we will discuss later. One may also consider a penalty that is additive in \mathbf{u}_k and \mathbf{v}_k :

$$P(\lambda_k, (d_k, \mathbf{u}_k, \mathbf{v}_k)) = \lambda_{k1} \sum_{i=1}^m w_{1,ik} |d_k u_{ik}| + \lambda_{k2} \sum_{j=1}^n w_{2,jk} |d_k v_{jk}|, \quad (1.6)$$

where λ_{k1} and λ_{k2} are two different regularization parameters. The penalty in (1.6) allows different degrees of sparsity to be imposed on \mathbf{u}_k and \mathbf{v}_k . This flexibility comes at

the cost of introducing extra regularization parameter, which reduces the computation efficiency. Meanwhile, the penalty term in (1.5) uses only one regularization parameter. However, due to its multiplicative form, it actually penalizes each SVD layer entrywisely, which leads to automatic adjustment of the degrees of sparsity between \mathbf{u}_k and \mathbf{v}_k . Therefore nothing is lost in terms of identifying the true sparse structure. Here, we mainly focus on the penalty (1.5), and the methodology can be easily extended to the penalty (1.6).

The rest of the article is organized as follows. We develop the methodology for the unit rank case in Section 2. We then discuss the extension to higher rank cases in Section 3. Two applications and simulation studies illustrating our method are given in Section 4. Some asymptotic results of the proposed method are presented in Section 5. We then conclude in Section 6.

2 Sparse Unit-rank Regression

In this section, we present the details of fitting the penalized regression model in (1.4) with the penalty given as (1.5) when the true coefficient matrix \mathbf{C} is of unit rank. We then present the extension to higher rank cases in Section 3.

2.1 Optimization algorithm and Initial Values

The problem here is to minimize the following penalized sum-of-squares criterion with respect to the triplets $(d, \mathbf{u}, \mathbf{v})$,

$$\frac{1}{2} \text{tr}[(\mathbf{S} - d\mathbf{u}\mathbf{v}^T \mathbf{G})(\mathbf{S} - d\mathbf{u}\mathbf{v}^T \mathbf{G})^T] + \lambda \sum_{i=1}^m \sum_{j=1}^n w_{ij} |du_i v_j|. \quad (2.1)$$

where $d\mathbf{u}\mathbf{v}^T$ is the SVD of the coefficient matrix \mathbf{C} ; $\mathbf{W} = (w_{ij})_{m \times n}$ consists of possibly data driven weights.

Following Zou (2006), the weights can be chosen as

$$\mathbf{W} = |\tilde{\mathbf{C}}|^{-\gamma} = |\tilde{d}\tilde{\mathbf{u}}\tilde{\mathbf{v}}^T|^{-\gamma}$$

where $\tilde{d}\tilde{\mathbf{u}}\tilde{\mathbf{v}}^T$ is the SVD of a \sqrt{T} -consistent estimator $\tilde{\mathbf{C}}$ of \mathbf{C} , e.g. the classical reduced-rank least square estimator given in Section 5, and γ is a positive tuning number. Note that choosing $\gamma = 0$ corresponds to the lasso fit (Tibshirani, 1996). In the following, we let $w_d = |\tilde{d}|^{-\gamma}$, $\mathbf{w}_1 = (w_{1,1}, \dots, w_{1,m})^T = |\tilde{\mathbf{u}}|^{-\gamma}$, $\mathbf{w}_2 = (w_{2,1}, \dots, w_{2,n})^T = |\tilde{\mathbf{v}}|^{-\gamma}$, $\mathbf{W}_1 = \text{diag}(\mathbf{w}_1)$ and $\mathbf{W}_2 = \text{diag}(\mathbf{w}_2)$ be given.

The model admits a biconvex structure in \mathbf{u} and \mathbf{v} . For fixed \mathbf{v} , minimization of (2.1) with respect to (d, \mathbf{u}) becomes minimization with respect to $\tilde{\mathbf{u}} = d\mathbf{W}_1\mathbf{u}$ of

$$\frac{1}{2} \|\mathbf{y} - \mathbf{X}_{(\mathbf{v})} \tilde{\mathbf{u}}\|_2^2 + \lambda_{(\mathbf{v})} \sum_{i=1}^m |\tilde{u}_i| \quad (2.2)$$

where $\mathbf{y} = \text{vec}(\mathbf{S}^T)$, $\mathbf{X}_{(\mathbf{v})} = \mathbf{W}_1^{-1} \otimes (\mathbf{G}^T \mathbf{v})$, and $\lambda_{(\mathbf{v})} = \lambda w_d(\sum_{j=1}^n w_{2,j} |v_j|)$. This can be recognized as a lasso regression problem with respect to $\check{\mathbf{u}}$, without an intercept term.

On the other hand, for fixed \mathbf{u} , minimization of (2.1) with respect to (d, \mathbf{v}) becomes minimization with respect to $\check{\mathbf{v}} = d\mathbf{W}_2\mathbf{v}$ of

$$\frac{1}{2} \|\mathbf{y} - \mathbf{X}_{(\mathbf{u})}\check{\mathbf{v}}\|_2^2 + \lambda_{(\mathbf{u})} \sum_{j=1}^n |\check{v}_j| \quad (2.3)$$

where $\mathbf{X}_{(\mathbf{u})} = \mathbf{u} \otimes \mathbf{G}^T \mathbf{W}_2^{-1}$, $\lambda_{(\mathbf{u})} = \lambda w_d(\sum_{i=1}^m w_{1,i} |u_i|)$, and \mathbf{y} is defined as above. Again, this is a lasso regression problem with respect to $\check{\mathbf{v}}$, without an intercept term.

We can take advantage of the biconvex structure of the objective function (2.1) in optimization. Here are the steps of our numerical algorithm for a fixed λ :

Sparse Unit-rank Regression Algorithm

1. Choose an initial value for $\check{\mathbf{v}}$.
2. Given fixed \mathbf{v} , solve lasso problem (2.2) to get $\check{\mathbf{u}}$ by either LARS algorithm (Efron et al., 2004) or coordinate descent algorithm (Friedman, 2007). Update $\hat{\mathbf{u}}$ and \hat{d} by normalizing $\mathbf{W}_1^{-1}\check{\mathbf{u}}$.
3. Given fixed \mathbf{u} , solve lasso problem (2.3) to get $\check{\mathbf{v}}$ by either LARS algorithm or coordinate descent algorithm. Update $\hat{\mathbf{v}}$ and \hat{d} by normalizing $\mathbf{W}_2^{-1}\check{\mathbf{v}}$.
4. Repeat steps 2-3, until $\hat{\mathbf{u}}\hat{d}\hat{\mathbf{v}}^T$ converges according to some stopping criterion.

The algorithm described above uses a block coordinate 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 model so that the existing fast algorithms for the lasso can be directly applied. It is clear that the criterion function is monotone decreasing along the iterations. The algorithm is therefore stable and guaranteed to converge, although not necessarily to the global minimum of the objective function. The biconvex optimization problems may have multiple local minima as in general they are global optimization problems, which requires more complicated algorithms to solve (Gorski et al., 2007). Nevertheless, we have not observed this to be a significant problem here.

The estimated coefficients vary with λ and produce a path of solutions regularized by λ . Based on numerical experiments, it appears that the solution paths for the above methods are continuous, but are not piecewise linear, unlike those for the lasso. In practice, the range of λ values one is interested in equals $[0, \lambda_{\max}]$, where λ_{\max} is the value at which all penalized coefficients are zero. Because the paths are continuous, a reasonable approach of choosing initial values is to start at λ_{\max} and use the estimate from the previous value of λ as the initial value for the next value of λ . The following lemma determines λ_{\max} .

Lemma 1. Denote $\mathbf{S} = [\mathbf{s}_{(1)}, \dots, \mathbf{s}_{(m)}]^T$ and $\mathbf{G} = [\mathbf{g}_{(1)}, \dots, \mathbf{g}_{(n)}]^T$. Then

$$\lambda_{max} = \max\left\{ \left| \frac{1}{w_{ij}} \mathbf{s}_{(i)}^T \mathbf{g}_{(j)} \right|, i = 1, \dots, m; j = 1, \dots, n. \right\}$$

Moreover, let $(i^*, j^*) = \operatorname{argmax}_{(i,j)} \left| \frac{1}{w_{ij}} \mathbf{s}_{(i)}^T \mathbf{g}_{(j)} \right|$, then the last nonzero solution of (2.1) denoted as $(\mathbf{u}^{(0)}, \mathbf{v}^{(0)})$ is given by

$$\begin{aligned} u_{i^*}^{(0)} &= 1; & u_i^{(0)} &= 0, i \neq i^*; \\ v_{j^*}^{(0)} &= 1; & v_j^{(0)} &= 0, j \neq j^*. \end{aligned}$$

Proof: The minimization problem (2.1) has the same λ_{max} as the lasso regression model

$$\frac{1}{2} \|\mathbf{y} - \mathbf{H}\rho\|_2^2 + \lambda w_d \sum_{i=1}^{mn} |\rho_i|$$

where $\mathbf{y} = \operatorname{vec}(\mathbf{S}^T)$, $\mathbf{H} = \mathbf{W}_1^{-1} \otimes (\mathbf{G}^T \mathbf{W}_2^{-1})$ and $\rho = (\rho_1, \dots, \rho_{mn})^T$ is an $mn \times 1$ vector. Then λ_{max} can be obtained as above by the KKT condition for lasso.

To find the solution path, we can start at $\lambda_{max} - \epsilon$ using $(\mathbf{u}^{(0)}, \mathbf{v}^{(0)})$ as initial values where ϵ is a very small positive number, and proceed towards 0 or to a minimum value λ_{min} at which the model becomes excessively large or ceases to be identifiable. This approach works very well in practice, and the algorithm usually converges within only a few iterations. Note that the reversed approach, which starts from small λ and goes toward large λ , may fail sometimes. This is because when λ is large, without a carefully chosen initial value, the inner updating step could produce zero solution for either left or right singular vector so that the algorithm can not proceed.

For all the numerical results in this paper, 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 fit, it is important to be able to choose an optimal point along the path. One common method that is used in practice is cross-validation (CR), based on the predictive performance of models (Stone, 1974). For small-scale problems, the optimal λ can be chosen by leave-one-out cross-validation, or more generally, K -fold cross-validation. However, in large-scale problems the CR method is usually very expensive from a computational point of view. Here we also consider three

widely used information criteria:

$$\begin{aligned}
AIC(\lambda) &= \log(SSE) + \frac{2}{mT} df_\lambda \\
BIC(\lambda) &= \log(SSE) + \frac{\log(mT)}{mT} df_\lambda \\
GCV(\lambda) &= \frac{SSE}{mn - df_\lambda}
\end{aligned} \tag{3.4}$$

where SSE stands for the sum of square errors, and df_λ is the effective number of parameters. The optimal value of λ is chosen to be the one that minimizes ~~the~~ criterion.

Zou et al. (2007) showed that the number of nonzero coefficients is an unbiased estimate for the degrees of freedom of the lasso problem. The biconvex objective function we consider here admits a conditional lasso structure, as shown in previous section. Therefore, we propose the following estimator for df_λ . Let $(\hat{d}^{(\lambda)}, \hat{\mathbf{u}}^{(\lambda)}, \hat{\mathbf{v}}^{(\lambda)})$ denote the fitted value of $(d, \mathbf{u}, \mathbf{v})$, then

$$\hat{df}_\lambda = \sum_{i=1}^m I(\hat{u}_i^{(\lambda)} \neq 0) + \sum_{j=1}^n I(\hat{v}_j^{(\lambda)} \neq 0)$$

where $I(\cdot)$ is the indicator function. We examine the performance of the proposed criteria via simulation in Section 4.

2.3 Orthogonal Design

To understand further the statistical properties of the proposed method, we consider the special case of orthogonal design. In fact there are many practical applications for this setting. ~~Especially~~, when ~~the~~ data matrix \mathbf{G} is an identity matrix, the regression model reduces to a biclustering problem (Busygin et al., 2008; Lee et al., 2010), which seeks simultaneous clustering of the rows and columns of a data matrix \mathbf{S} . It is of great interest to be able to obtain a sparse decomposition of the matrix so that interpretable row-column associations can be identified. Without loss of generality, in the following we consider $\mathbf{G}\mathbf{G}^T = I$, i.e. \mathbf{G} is orthonormal.

The following lemma gives a necessary condition for the minimizer of expression (2.1) in ~~the~~ orthogonal design situation.

Lemma 2. Suppose $\mathbf{G}\mathbf{G}^T = I$. Then the solution $(\hat{d}, \hat{\mathbf{u}}, \hat{\mathbf{v}})$ of model (2.1) satisfies ~~the~~

$$\begin{aligned}
\hat{d}\hat{\mathbf{u}} &= sign(\tilde{\mathbf{u}}^*)(|\tilde{\mathbf{u}}^*| - \lambda_{(\hat{\mathbf{v}})}\mathbf{w}_1)^+ \\
\hat{d}\hat{\mathbf{v}} &= sign(\tilde{\mathbf{v}}^*)(|\tilde{\mathbf{v}}^*| - \lambda_{(\hat{\mathbf{u}})}\mathbf{w}_2)^+
\end{aligned}$$

where

$$\begin{aligned}\tilde{\mathbf{u}}^* &= \mathbf{S}\mathbf{G}^T\hat{\mathbf{v}} = (\hat{\mathbf{v}}^T\mathbf{G}\mathbf{s}_{(1)}, \dots, \hat{\mathbf{v}}^T\mathbf{G}\mathbf{s}_{(m)})^T; \\ \tilde{\mathbf{v}}^* &= \mathbf{G}\mathbf{S}^T\hat{\mathbf{u}} = (\hat{\mathbf{u}}^T\mathbf{S}\mathbf{g}_{(1)}, \dots, \hat{\mathbf{u}}^T\mathbf{S}\mathbf{g}_{(n)})^T; \\ \lambda_{(\hat{\mathbf{u}})} &= \lambda w_d \sum_{i=1}^m w_{1,i} |\hat{u}_i|; \\ \lambda_{(\hat{\mathbf{v}})} &= \lambda w_d \sum_{j=1}^n w_{2,j} |\hat{v}_j|.\end{aligned}$$

Lemma 2 shows that when either \mathbf{u} or \mathbf{v} is fixed, the other one along with the singular value d can be estimated via a simple soft-thresholding rule. The lemma leads to a more efficient computation algorithm based on coordinate descend method, which can be quite demanding for high dimensional problems. 

3 Extension to Higher Ranks

As stated previously, we assume the rank of \mathbf{C} has been correctly identified to be r . We present here several different ways to extend the unit-rank methodology to the higher rank cases.

In order to obtain sparse estimates of multiple layers, one naive approach is to minimize the unit-rank criterion (2.1) repeatedly, each time using as the \mathbf{S} matrix the residuals obtained by subtracting from the data matrix the previous layers found. The algorithm is as follows,

Sequential-extraction Algorithm

1. Let $\mathbf{S}_1 = \mathbf{S}$.
2. For $k \in 1, \dots, r$:
 - (1) Find $(\hat{\mathbf{u}}_k, \hat{d}_k, \hat{\mathbf{v}}_k)$ by performing the sparse unit-rank regression of \mathbf{S}_k on \mathbf{G} . The optimal λ_k is the minimizer of some criterion defined in (3.4), e.g. $BIC(\lambda_k)$.
 - (2) Let $\mathbf{S}_{k+1} = \mathbf{S}_k - \hat{d}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$
3. The final estimate of \mathbf{C} is given by $\hat{\mathbf{C}} = \sum_{k=1}^r \hat{d}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$.

This sequential fitting idea has been used in many penalized matrix decomposition methods (Witten et al., 2009; Lee et al., 2010) and it generally works well. Without the penalty constraints, it can be shown that the sequential-extraction algorithm leads to the rank- r reduced-rank regression of \mathbf{S} on \mathbf{G} . In particular, the successive solutions are orthogonal. With the penalty presents, the orthogonality property does not hold any more. We find by simulation that when the singular values are close to each other,

the algorithm may fail to distinguish between different layers. As a consequence, it may produce incorrect associations between the responses and the covariates.

Now suppose some \sqrt{T} -consistent estimate $\tilde{\mathbf{C}}$ of \mathbf{C} , say, $\tilde{\mathbf{C}}$ is available, whose SVD is given by $\sum_{k=1}^r \tilde{d}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^T$. Note that this implies $\tilde{\mathbf{u}}_k$ s, \tilde{d}_k s and $\tilde{\mathbf{v}}_k$ s are all \sqrt{T} -consistent for \mathbf{u}_k s, d_k s and \mathbf{v}_k s, respectively by Lemma 3. ~~Usually the initial estimate can be easily obtained by the classical reduced-rank regression. For high dimensional data, to avoid the SVD of a $m \times m$ matrix, one can obtain the reduced-rank estimate of \mathbf{C} by performing the above sequential-extraction algorithm without the penalty term. We then propose to obtain the sparse estimates of \mathbf{C} by the following exclusive-extraction algorithm:~~

Exclusive-extraction Algorithm

1. For $k \in 1, \dots, r$:
 - (1) Let $\mathbf{S}_k = \mathbf{S} - \tilde{\mathbf{C}}_{-k} \mathbf{G}$ with $\tilde{\mathbf{C}}_{-k} = \tilde{\mathbf{C}} - \tilde{\mathbf{C}}_k$. Let $\mathbf{W}_k = |\tilde{\mathbf{C}}_k|^{-\gamma} = |\tilde{\mathbf{u}}_k \tilde{d}_k \tilde{\mathbf{v}}_k^T|^{-\gamma}$, where $\tilde{\mathbf{u}}_k \tilde{d}_k \tilde{\mathbf{v}}_k^T$ is the SVD of the layer- k estimate $\tilde{\mathbf{C}}_k$.
 - (2) Find $(\hat{\mathbf{u}}_k, \hat{d}_k, \hat{\mathbf{v}}_k)$ by performing the sparse unit-rank regression of \mathbf{S}_k on \mathbf{G} . The optimal λ_k is the minimizer of some criterion defined in (3.4), e.g. $BIC(\lambda_k)$.
2. The final estimate of \mathbf{C} is given by $\hat{\mathbf{C}} = \sum_{k=1}^r \hat{d}_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T$.

The above method seeks the sparse estimator of $\hat{\mathbf{C}}$ by separately solving r sparse unit-rank regression problems. The computation cost increases linearly as the rank r increases, and the estimation for different layers can be performed in parallel. The simulation results confirm that this algorithm works better than the sequential-extraction method in general. We have also proved that the estimator obtained by exclusive-extraction method enjoys many good large sample properties. In practice, the quality of the estimation may partly depend on the initial estimator of \mathbf{C} which is used to form the exclusive layers. Especially, when the dimension is high relative to the sample size and the true model is very sparse, the classical reduced-rank regression estimate $\tilde{\mathbf{C}}$ might not be a good choice for forming the exclusive layers.

Note that we have shown that the estimates obtained by the exclusive-extraction method ~~satisfy the \sqrt{T} -consistency condition~~. To further improve the estimation, we propose the following iterative exclusive-extraction algorithm:

Iterative Exclusive-extraction Algorithm

1. Use the initial estimate $\tilde{\mathbf{C}}$ to form the exclusive layers, and use the exclusive-extraction algorithm to get the sparse estimate $\hat{\mathbf{C}}^{(1)}$.
2. Use the estimate $\hat{\mathbf{C}}^{(i)}$ to form the exclusive layers, and use the exclusive-extraction algorithm to get the sparse estimate $\hat{\mathbf{C}}^{(i+1)}$.

3. Repeat step 2 for some times, or until $\hat{\mathbf{C}}^{(i)}$ converges according to some stopping criteria.



Interestingly, the above iterative algorithm can be regarded as a way to solve the general optimization problem (1.4). It has a block coordinate decent structure, with each SVD layer as one block. The selection of the regularization parameters is nested within the iterative algorithm, which prevents using the computationally more expensive simultaneous selection of η_1 parameters. Our experience from simulation studies and real applications suggests that the iterative algorithm typically converges within only a few iterations.

4 Simulation and Real Applications

4.1 Simulation 1: Unit-rank Biclustering

In this simulation, we consider a unit-rank biclustering problem. Let \mathbf{G} be a 50×50 identity matrix, and let $\mathbf{S}^* = d\mathbf{u}\mathbf{v}^T$ be a 100×50 unit-rank matrix with $d = 50$ and

$$\check{\mathbf{u}} = [10, 9, 8, 7, 6, 5, 4, 3, \text{rep}(2, 17), \text{rep}(0, 75)]^T, \quad \mathbf{u} = \check{\mathbf{u}}/\|\check{\mathbf{u}}\|_2; \\ \check{\mathbf{v}} = [10, -10, 8, -8, 5, -5, \text{rep}(3, 5), \text{rep}(-3, 5), \text{rep}(0, 34)]^T, \quad \mathbf{v} = \check{\mathbf{v}}/\|\check{\mathbf{v}}\|_2$$

where $\text{rep}(a, b)$ denotes a vector of length b , whose entries are all a. A data matrix \mathbf{S} is generated as the sum of \mathbf{S}^* and the noise matrix \mathbf{E} , whose elements are randomly sampled from the standard normal distribution, which makes the signal to noise ratio (SNR) approximately equals to 0.5. The nonzero entries of \mathbf{S}^* take on several distinct values, some of which are quite small. This makes the model estimation very challenging. We use rank-1 approximation of \mathbf{S} based on SVD and $\gamma = 2$ in deciding the adaptive weights \mathbf{w}_1 and \mathbf{w}_2 , and the optimal solution along the path is chosen based on BIC.

		Avg. # of zeros (true)	Avg. # of correctly identified zeros	Avg. # of correctly identified nonzeros	Misclassification rate
SRRR	\mathbf{u}	73.89(75)	73.65(98.20%)	24.76(99.04%)	1.06%
	\mathbf{v}	33.90(34)	33.90(99.70%)	16.00(100.0%)	0.07%
SSVD	\mathbf{u}	74.33(75)	74.03(98.70%)	24.70(98.78%)	0.85%
	\mathbf{v}	33.79(34)	33.79(99.39%)	16.00(100.0%)	0.14%

Table 1: Simulation 1: Comparison of the performance between SRRR and SSVD.

Lee et al. (2010) used this example to compare their sparse singular value decomposition (SSVD) method with several other popular biclustering methods. The SSVD

method used the penalty (1.6) with \mathbf{G} being an identity matrix. Therefore it can be regarded as a special case of our method. The regularization parameter selection was nested within the coordinate decent iterations. However, Lee et al. (2010) did not study the theoretical properties of their estimator. Nevertheless, Lee et al. (2010) found that the SSVD method performed much better than the other methods in terms of having lower misclassification rate. Therefore here we only compare our method with the SSVD method. The simulation is repeated 1000 times. Table 1 reports the simulation results of our method and of the SSVD method for comparison. ~~As one can see, the misclassification rates given by our method are also extremely low, which are comparable with the SSVD method. Note that our method uses the grid search strategy to select the regularization parameter, which is more standard and of course is of higher computational cost. However, since only one regularization parameter is used and we have taken advantage of the continuity property of the solution path, the computation of our method is extremely fast. The whole simulation only takes a few seconds.~~

4.2 Simulation 2: Mimic the ecological application

In this example, we try to mimic the ecological application discussed in the next section. Let \mathbf{G} be a 45×27 matrix, whose columns are independently generated from $AR(1)$ process, ~~for which the auto-regressive parameter is 0.4 and the error standard deviation is 1~~. Let $\mathbf{C} = d\mathbf{u}\mathbf{v}^T$ be a 18×45 unit-rank matrix with $d = 1$ and

$$\check{\mathbf{u}} = [rep(1, 9), rep(0, 9)]^T, \quad \mathbf{u} = \check{\mathbf{u}}/\|\check{\mathbf{u}}\|_2; \\ \check{\mathbf{v}} = [v_1, \dots, v_{45}]^T, \text{ with } v_j = (j-1)^4(j-45)^2, \quad \mathbf{v} = \check{\mathbf{v}}/\|\check{\mathbf{v}}\|_2.$$

The \mathbf{v} vector forms a polynomial curve which peaks at the thirtieth position, see Figure 1. The underlying assumption here is that the true spawning curve is smooth over the 45-day period and the spawning activity reaches its highest level at the thirtieth day. A data matrix \mathbf{S} is generated as the sum of \mathbf{CG} and the noise matrix \mathbf{E} , whose elements are ~~randomly sampled from the normal distribution with mean 0 and standard deviation σ~~ . In each replication, σ is chosen so that the SNR is of a certain level. We use the same fitting procedure as in the actual ecological application. The simulation is repeated 100 times.

Three different regularization parameter selection methods are used, namely, the BIC, the AIC, and the GCV. In Table 2, we report the false discovery rate (FDR) and the false negative rate (FNR) based on $\hat{\mathbf{u}}$. It can be seen that our method works reasonably well in term of ~~the fjord selection~~. Both FDRs and FNRs decreases as the signal to noise ratio increases. The BIC yields the smallest FDRs among the three criterion. However, the FNRs of the BIC is generally higher than ~~the FNRs~~ of the AIC or the GCV, especially when the signal to noise ratio is low.

Due to the fusion-type penalty imposed on the right singular vector in our model, the estimated spawning curve $\hat{\mathbf{v}}$ can successfully capture the general form of the true curve, i.e. the estimated curve is piecewise linear, equals zero at both ends and peaks at

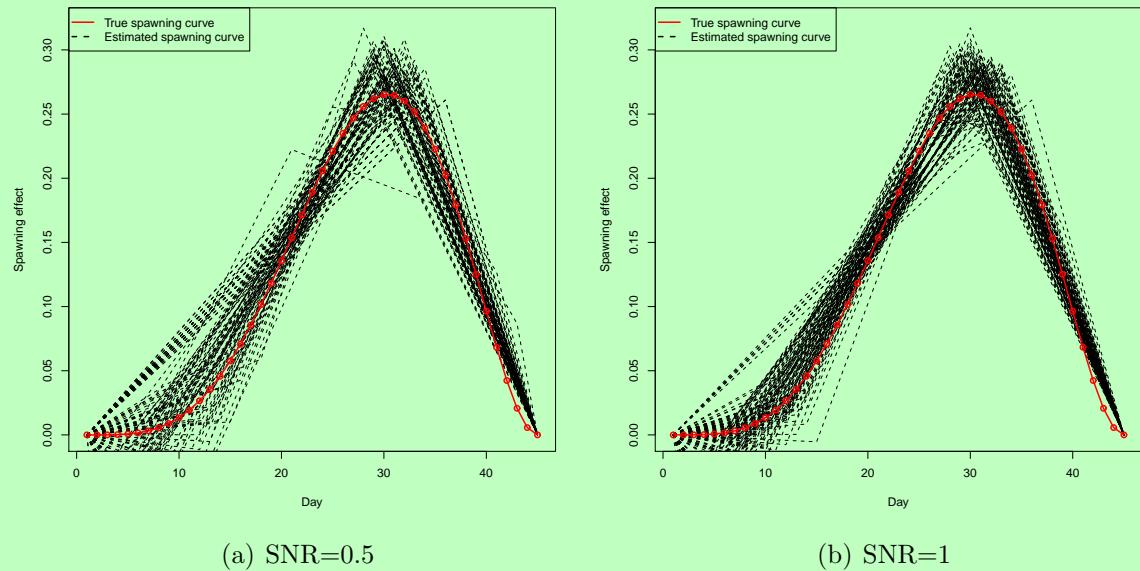


Figure 1: Simulation 2: Estimated spawning curves (black) and the true spawning curve (red). Left: SNR=0.5; Right: SNR=1.

a position right at or near the true peak, see Figure 1. Also reported in Table 2 are the averaged percentile of \hat{v}_{30} (APV), and the averaged distance from the estimated peak to the true peak (ADV). Both the APV and the ADV measure whether the selected peak is close to the position of the true peak. It can be seen from the results that the estimated peak is very close to the true peak most of the time. Even when the signal to noise ratio is very low, the estimated peak is not far off.

4.3 Simulation 3: Higher Rank Example

In this simulation, we let \mathbf{G} be a 30×50 matrix whose entries are independently generated from the standard normal distribution, and let C be a 30×100 rank-3 matrix

Criterion	Rates	Signal to noise ratio (SNR)					
		0.0625	0.125	0.25	0.5	1	2
BIC	FDR	8.49%	5.59%	7.69%	5.30%	1.68%	0.20%
	FNR	80.44%	42.33%	4.00%	0.11%	0.00%	0.00%
	APV	77.87%	88.47%	91.00%	94.73%	95.38%	95.84%
	ADV	6.75	3.51	3.01	1.83	1.58	1.48
AIC	FDR	32.90%	22.76%	14.39%	6.08%	1.86%	0.20%
	FNR	25.89%	7.67%	1.89%	0.00%	0.00%	0.00%
	APV	85.58%	89.76%	93.16%	94.60%	95.20%	95.64%
	ADV	6.17	3.67	2.33	1.87	1.75	1.61
GCV	FDR	36.82%	24.78%	15.86%	6.56%	1.86%	0.20%
	FNR	18.33%	5.89%	1.44%	0.00%	0.00%	0.00%
	APV	84.76%	89.47%	93.02%	94.44%	95.18%	95.47%
	ADV	7.28	3.86	2.44	2.00	1.81	1.70

Table 2: Simulation 2: Simulation results for mimicking the ecological application. FDR: false discovery rates of \mathbf{u} ; FNR: false negative rate of \mathbf{u} ; APV: averaged percentile value of \hat{v}_{30} ; ADV: averaged distance from the estimated peak to the true peak.

whose SVD is given by $\sum_{k=1}^3 d_k \mathbf{u}_k \mathbf{v}_k^T$ with

$$\begin{aligned}
\check{\mathbf{u}}_1 &= [sample(\pm 1, 15), rep(0, 85)]^T; \\
\check{\mathbf{u}}_2 &= [rep(0, 20), sample(\pm 1, 15), rep(0, 65)]^T; \\
\check{\mathbf{u}}_3 &= [rep(0, 9), \check{\mathbf{u}}_{1,10:12}, -\check{\mathbf{u}}_{1,13:15}, sample(\pm 1, 5), -\check{\mathbf{u}}_{2,21:22}, \check{\mathbf{u}}_{2,23:24}, rep(0, 76)]^T; \\
\check{\mathbf{v}}_1 &= [unif(-0.5, 0.5, 10), rep(0, 20)]^T; \\
\check{\mathbf{v}}_2 &= [rep(0, 10), sample(\pm 1, 10) * unif(0.5, 1, 10), rep(0, 10)]^T; \\
\check{\mathbf{v}}_3 &= [rep(0, 20), sample(\pm 1, 10) * unif(0.5, 1, 10)]^T; \\
\mathbf{u}_k &= \check{\mathbf{u}}_k / \|\check{\mathbf{u}}_k\|_2, \mathbf{v}_k = \check{\mathbf{v}}_k / \|\check{\mathbf{v}}_k\|_2 \text{ for } k = 1, 2, 3; \\
d_1 &= 15, d_2 = 10, d_3 = 5.
\end{aligned}$$

where $sample(\mathcal{A}, b)$ denotes a vector of length b , whose entries are independently sampled from \mathcal{A} with replacement, and $unif(a_1, a_2, b)$ also denotes a vector of length b , whose entries are independently sampled from a $Uniform(a_1, a_2)$ distribution. A data matrix \mathbf{S} is generated as the sum of \mathbf{CG} and the noise matrix \mathbf{E} , whose elements are randomly sampled from the normal distribution with mean 0 and standard deviation σ . In each replication, σ is chosen so that the signal to noise ratio calculated based on the third layer $\mathbf{u}_3 d_3 \mathbf{v}_3^T$ and the noise matrix \mathbf{E} is of a certain level. The nonzero entries of \mathbf{u}_k s have some positional overlap with each other, and the nonzero entries of \mathbf{v}_k s take on distinct values, some of which are quite small. Moreover, the singular values

are quite close to each other. These make the model estimation more challenging.

We consider several methods, i.e. the exclusive-extraction method, the iterative exclusive-extraction method, and the sequential-extraction method. For the iterative exclusive-extraction method, we only run one additional iteration. For the sequential-extraction method, we consider two ways of obtaining the adaptive weights. One way is to use coefficient estimates from sequentially performing unit-rank regression of the previous residual data matrix. Another way is to use coefficient estimates from an initial rank-3 regression. The simulation is repeated 100 times for each signal to noise ratio. The optimal solution along the path is chosen based on BIC, and we use $\gamma = 2$ in ~~deciding~~ the adaptive weights.

Table 3 reports the estimation results for comparison. Overall the iterative exclusive-extraction method works the best in terms of having the lowest FDR and well-controlled FNR. Not surprisingly, the sequential-extraction method with sequential weights works the worst. Its FDR is much higher than the FDRs of the other methods due to its incapability of distinguishing the different layers ~~sometimes~~, and its FDR does not seem to decrease as the SNR increases. It is interesting to see that using the weights constructed from ~~an~~ initial rank-3 regression can improve the sequential fitting a lot.

4.4 Modeling Larval Drift Effects on Cod Population Dynamics

In Norway, a beach-seine monitoring program was begun in the early 1900s to collect data on fall abundance of 6-month old fish in several fjords along the Norwegian Skagerrak coast, which is still going on. Chan et al. (2003a) developed a fjord-based ARMAX(2,2) time series model using the beach-seine data for ~~studying~~ the cod population dynamics. The model considered a series of ~~coastal locations (or fjords, see in~~ Figure 2) to represent demographically (semi-) autonomous populations. It incorporated within- and between-cohort interactions, interactions with coexisting species, and several environmental factors. Stenseth et al. (2006) ~~applied~~ the ARMAX(2,2) model to evaluate the hypothesis that Atlantic cod larvae are passively transported by sea currents from ~~off shore spawning areas to settle in the Norwegian Skagerrak waters.~~ This finding for the first time demonstrated a direct link between larval drift and gene flow ~~in the Skagerrak marine environment.~~ Here our objective is to ~~further evaluate~~ the hypothesis that ~~the cod population dynamics within~~ a certain coastal fjord ~~may depend~~ on the fjord's potential of receiving the North Sea larvae.

We analyze the same 15 fjords studied in Chan et al. (2003a), Chan et al. (2003b) and Stenseth et al. (2006). The beach-seine stations within these 15 fjords are classified and recombined into 9 exposed fjords and 9 inner fjords based on ~~the evaluation about~~ their degree of exposure ~~to the larval drift from external sources and their geographical proximity.~~ The logarithmically transformed time series of 0-group (i.e., fish that are ~~0-6~~ months old) cod abundance of each fjord (exposed or inner) are calculated following similar weighting scheme as used in Chan et al. (2003a) and Chan et al. (2003b). We thus first fit the fjord-based ARMAX(2,2) population dynamics model for the 9 exposed

Methods	Rates	Layers	Signal to noise ratio					
			0.03125	0.0625	0.125	0.25	0.5	1
Exclusive(1)	FDR	Layer 1	5.57%	3.81%	3.03%	2.37%	1.81%	0.94%
		Layer 2	10.13%	3.98%	2.19%	1.67%	1.54%	1.85%
		Layer 3	10.80%	11.15%	5.05%	1.90%	1.05%	1.00%
		Overall	9.04%	6.66%	3.61%	2.06%	1.54%	1.34%
	FNR	Layer 1	7.72%	4.76%	3.52%	2.88%	1.92%	1.48%
		Layer 2	3.76%	0.08%	0.00%	0.00%	0.00%	0.00%
		Layer 3	74.84%	7.96%	0.88%	0.08%	0.00%	0.00%
		Overall	28.77%	4.27%	1.47%	0.99%	0.64%	0.49%
Exclusive(2)	FDR	Layer 1	3.23%	1.91%	1.43%	0.55%	0.62%	0.16%
		Layer 2	6.09%	2.29%	1.00%	0.34%	0.61%	0.45%
		Layer 3	6.84%	8.96%	4.11%	1.44%	0.78%	0.55%
		Overall	5.39%	4.57%	2.29%	0.81%	0.69%	0.40%
	FNR	Layer 1	7.48%	4.92%	3.64%	3.04%	2.04%	1.40%
		Layer 2	3.36%	0.04%	0.00%	0.00%	0.00%	0.00%
		Layer 3	80.20%	8.04%	0.92%	0.08%	0.00%	0.00%
		Overall	30.35%	4.33%	1.52%	1.04%	0.68%	0.47%
Sequential(1)	FDR	Layer 1	4.22%	6.42%	6.81%	8.95%	15.52%	13.62%
		Layer 2	4.79%	3.78%	5.06%	6.83%	13.19%	14.39%
		Layer 3	3.43%	8.08%	4.09%	2.01%	4.16%	8.18%
		Overall	4.67%	6.50%	5.91%	6.49%	12.30%	13.27%
	FNR	Layer 1	8.28%	5.92%	4.80%	5.72%	3.80%	3.28%
		Layer 2	1.16%	0.16%	0.00%	2.00%	0.00%	0.00%
		Layer 3	80.40%	8.72%	1.00%	0.04%	0.00%	0.00%
		Overall	29.95%	4.93%	1.93%	2.59%	1.27%	1.09%
Sequential(2)	FDR	Layer 1	2.60%	1.53%	1.13%	1.50%	2.59%	1.62%
		Layer 2	5.48%	1.64%	0.99%	0.54%	0.48%	1.06%
		Layer 3	6.32%	8.93%	4.10%	1.34%	0.91%	1.64%
		Overall	4.76%	4.22%	2.19%	1.21%	1.45%	1.54%
	FNR	Layer 1	7.64%	5.24%	4.16%	3.80%	3.36%	2.76%
		Layer 2	3.32%	0.04%	0.00%	0.00%	0.00%	0.00%
		Layer 3	82.36%	8.72%	0.88%	0.04%	0.00%	0.00%
		Overall	31.11%	4.67%	1.68%	1.28%	1.12%	0.92%

Table 3: Simulation 3: Comparison of the performance between exclusive-extraction method and sequential-extraction method. FDR: false discovery rate; FNR: false negative rate.

fjords and 9 inner fjords, and then analyze the residuals of the model. It is expected that part of the variation among the residuals can possibly be explained by the larvae drift ~~phenomenon~~. Here we let \mathbf{S} denote the $m \times T$ data matrix with $m = 18$ and $T = 27$, whose entry s_{it} is the residual for fjord i and year t of the ARMAX(2,2) model.

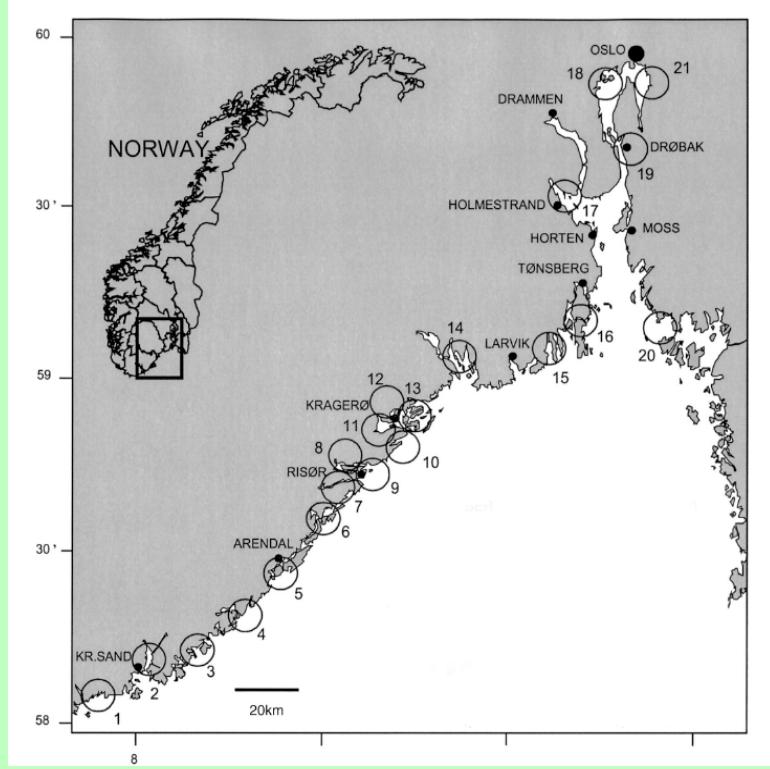


Figure 2: The Norwegian Skagerrak coastal area showing the 21 fjords where beach seine surveys have been conducted during the period from 1897 to 2008.

Cod larvae may potentially reach Skagerrak by passive current drift from the spawning ground in the North Sea. To quantify the amount of larvae drift, we first estimate the annual spawning biomass distribution over the North Sea from the International Bottom Trawl Survey (IBTS) data. Secondly, a oceanographic model is used to estimate the probability of larvae drift from the North Sea to Skagerrak, at a certain date and at a certain geographical grid on the spawning area. Finally, we obtain a proxy of daily larvae drift to Skagerrak as a weighted average of the spawning biomass over the North Sea spawning area with the drift probabilities being the weights, for the period from February 22nd to April 7th, a 45-day period that covers the potential spawning window, from year 1981 to 2007 for which the CPUE data are available. Here we let \mathbf{G} be a $n \times T$ matrix with $n = 45$ and $T = 27$, whose entry g_{jt} is the logarithmically transformed North Sea larvae drift proxy at day j of year t , for $j = 1, \dots, n$ and $t = 1, \dots, T$.

To study the larvae drift effects among the 18 coastal fjords, we propose the following

model 

$$s_{it} = u_{if} \sum_{j=1}^n v_j g_{jt} + e_{it} \quad i = 1, \dots, m; j = 1, \dots, n, \quad (4.1)$$

where u_i s can be regarded as the fjord effects, v_j s the daily spawning effects, and e_{it} s are assumed to be independently and identically distributed as $N(0, \sigma^2)$. Here without loss of generality, we have adjusted the response variables to be centered, i.e. $\sum_{t=1}^T s_{it} = 0$, for $i = 1, \dots, m$, and the predictors be standardized, i.e. $\sum_{t=1}^T g_{jt} = 0$, $\frac{1}{T} \sum_{t=1}^T g_{jt}^2 = 1$, for $j = 1, \dots, n$. Hence we do not need an intercept in the model.

Let $\mathbf{u} = (u_1, \dots, u_m)^T$, $\mathbf{v} = (v_1, \dots, v_n)^T$, and $\mathbf{E} = (e_{it})_{m \times T}$. To make the model identifiable, we restrict $\mathbf{u}^T \mathbf{u} = 1$ and $\mathbf{v}^T \mathbf{v} = 1$. Then (4.1) can be written in matrix form as

$$\mathbf{S} = d\mathbf{u}\mathbf{v}^T \mathbf{G} + \mathbf{E} \quad (4.2)$$

where d is a multiple, $\mathbf{u}^T \mathbf{u} = 1$ and $\mathbf{v}^T \mathbf{v} = 1$.

Model (4.2) can be recognized as a reduced-rank regression model with rank $r = 1$. In our study of the cod population dynamics here, the spawning curve (\mathbf{v}) should be a smoothed function of the day index, first increases and then decreases over the 45-day period. Moreover, the larvae drift effects among 18 fjords (\mathbf{u}) should be sparse, since we expect that only the exposed fjords can potentially receive larvae drift from external sources but not inner fjords. We then propose to estimate $(d, \mathbf{u}, \mathbf{v})$ by minimizing the following objective function:

$$\begin{aligned} & \frac{1}{2} \text{tr}[(\mathbf{S} - d\mathbf{u}\mathbf{v}^T \mathbf{G})(\mathbf{S} - d\mathbf{u}\mathbf{v}^T \mathbf{G})^T] \\ & + \lambda d \left[\sum_{i=1}^m w_{1,i} |u_i| \right] \left[w_{2,1} |v_1| + \sum_{j=2}^{n-1} w_{2,j} |(1-B)^2 v_{j+1}| + w_{2,n} |v_n| \right] \end{aligned}$$

where λ is the regularization parameter, B is the backshift operator and $w_{1,i}$ s and $w_{2,j}$ s are possibly data driven weights.

The penalty term in the above model admits a multiplicative form in \mathbf{u} and \mathbf{v} . The part about \mathbf{u} is a lasso-type penalty, which encourages sparsity in u_i s. The part about \mathbf{v} is a fusion-type penalty, which not only encourages smoothness in v_j s, but also forces \mathbf{v} to be small at its two ends. Another advantage of the fusion-type constraint on \mathbf{v} is that it can be easily transformed to the lasso constraint. Define $\theta_1 = v_1$, $\theta_j = (1-B)^2 v_{j+1}$ for $j = 2, \dots, n-1$ and $\theta_n = v_n$. Since the transformation from \mathbf{v} to $\theta \equiv (\theta_1, \dots, \theta_n)^T$ is one to one, there exists a unique $n \times n$ nonsingular matrix \mathbf{L} such that $\mathbf{v} = \mathbf{L}\theta$. Then we recognize that by re-defining \mathbf{G} to be $\mathbf{L}^T \mathbf{G}$ and \mathbf{v} to be θ , the above model has exactly the same form as the model (2.1). Hence we use the proposed sparse unit-rank regression method to carry out the above estimation problem. Note that several factors make the estimation very challenge. Firstly, the sample size T is 27, which is relatively small for $m = 18$ and $n = 45$. Secondly, the SNR is expected to be low. For regularization parameter selection, BIC tends to be too conservative when SNR is low based on our simulation study. Thus, more liberal model selection criteria

AIC and GCV are considered. We also use leave-one-out cross-validation (LOOCR), which involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data.

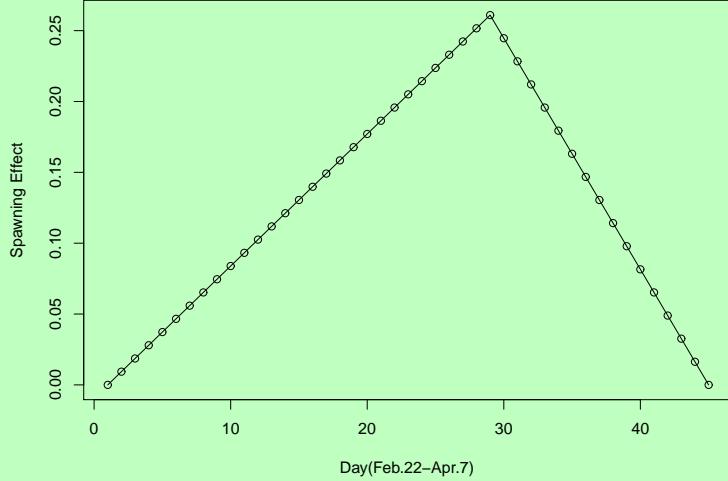


Figure 3: Estimated spawning effects (Feb.22-Apr.7).

All four criterion yield the same spawning curve estimate \hat{v} , see Figure 3. The estimated curve is piecewise linear and has a triangle shape, which is induced by the particular type of fusion penalty we have used in the model. The estimated peak is at the 29th day, which indicates that spawning in the North Sea peaks around mid-March, which is consistent with peer studies. The larvae drift effects are indeed sparse among all the fjords. In this case using BIC for the selection of λ results in a model including only one exposed fjord. Based on AIC, GCV and LOOCR, it is clear that the larvae drift effects are mostly seen among exposed fjords but not inner fjords. Among 9 inner fjords, only the inner fjords 10 and 19 appear to have strong larvae drift effects. Interestingly, the fjords 10 and 19 are among the only three fjords who have both inner and exposed parts, which suggests that the "inner" parts of these two fjords could have indirectly received contributions from North Sea cod through within-fjord migration of young cod. To better estimate the spawning peak, we drop the inner fjords 10 and 19, and re-fit the fusion-lasso model with the remaining 16 fjords. BIC, AIC and LOOCV yield the same spawning curve estimate as before, while the estimated curve based on GCV peaks at the 27th day, which is very close to the original estimate. Again, we find evidence that the larvae drift effects are mostly seen among exposed fjords but not inner fjords (Table 4.4).

Chen and Chan (2010) developed a bootstrap approach based on the ARMAX(2,2) model to test whether the cod population dynamics are common among the exposed

fjords and the inner Fjords. They found that ~~all~~ the cod dynamics are similar across the inner fjords, but the cod dynamics are different across the exposed fjords. Therefore, the finding in the current analysis suggests that the differential influence from North Sea cod larvae could be the cause of the dissimilarity ~~among~~ the cod dynamics of these exposed fjords.

Criterion	Fjord type	Fjord number/Larvae drift effect (\hat{u})								
AIC	Exposed	1	2	5	9	10	16	17	19	20
		0.00	0.38	0.00	0.28	0.47	0.07	0.00	0.46	0.00
	Inner	2	3	4	7	8	10	11	13	19
		0.00	0.00	0.00	0.00	0.00	0.34	0.00	0.00	0.47
GCV	Exposed	1	2	5	9	10	16	17	19	20
		0.22	0.33	0.00	0.27	0.38	0.13	0.22	0.47	0.00
	Inner	2	3	4	7	8	10	11	13	19
		-0.04	0.00	0.00	0.06	0.00	0.34	0.00	0.00	0.47
LOOCR	Exposed	1	2	5	9	10	16	17	19	20
		0.12	0.36	0.00	0.28	0.43	0.11	0.10	0.47	0.00
	Inner	2	3	4	7	8	10	11	13	19
		0.00	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.48

Table 4: Estimated larvae drift effects based on the 18-fjord model.

Criterion	Fjord type	Fjord number/Larvae drift effect (\hat{u})								
AIC	Exposed	1	2	5	9	10	16	17	19	20
		0.00	0.51	0.00	0.34	0.67	0.00	0.00	0.43	0.00
	Inner	2	3	4	7	8	11	13		
		0.00	0.00	0.00	0.00	0.00	0.00	0.00		
GCV	Exposed	1	2	5	9	10	16	17	19	20
		0.20	0.44	0.00	0.35	0.51	0.14	0.18	0.58	0.00
	Inner	2	3	4	7	8	11	13		
		0.00	0.00	0.00	0.00	0.00	0.00	0.00		
LOOCR	Exposed	1	2	5	9	10	16	17	19	20
		0.07	0.47	0.00	0.36	0.56	0.12	0.03	0.57	0.00
	Inner	2	3	4	7	8	11	13		
		0.00	0.00	0.00	0.00	0.00	0.00	0.00		

Table 5: Estimated larvae drift effects based on the 16-fjord model.

4.5 Biclustering: Lung Cancer Data

In this application, we illustrate by a real application the effectiveness of the proposed sparse reduced-rank regression method for microarray biclustering problem (Busyggin et al., 2008). The goal is to identify sets of biologically relevant genes that are significantly expressed for certain cancer types using microarray gene expression data, in which usually thousands of genes are measured for only a few subjects. The proposed method is well-suited for such a simultaneous selection problem. We show that our method is very flexible in that it can be either a unsupervised or supervised learn tool. Moreover, the method can be further extended to adjust the “unwanted” expression heterogeneity, so that a more complete statistical biclustering framework can be built upon.

The gene expression data consist of expression levels of $m = 12625$ genes, measured from $T = 56$ subjects. 17 subjects are known to be normal (Normal), and 39 patients are known to be with one of the three types of lung cancer. Among the patients, 20 of them are with pulmonary carcinoid tumors (Carcinoid), 13 of them are with colon metastases (Colon) and 6 of them are with small cell carcinoma (SmallCell). The data form a $m \times T$ matrix (\mathbf{S}) whose columns represent the subjects, grouped sequentially by the cancer type (Carcinoid, Colon, Normal and SmallCell), and the rows correspond to the genes. More detailed description of the data can be found in Bhattacharjee et al. (2001). A subset of the data was analyzed by Liu et al. (2008), in which they proposed a method called *SigClust* for assessing statistical significance of clusters. The data was also analyzed by Lee et al. (2010), in which the SSVD method was proposed for biclustering.

Suppose we let the covariate matrix \mathbf{G} be a $T \times T$ identity matrix, it can be seen that the sparse reduced-rank regression model actually reduces to a low-rank matrix approximation problem of the data matrix \mathbf{S} , with the sparsity requirement on the singular vectors. In this analysis, we use the iterative exclusive-extraction method. The computation is very fast due to the orthogonality of the \mathbf{G} matrix. We consider only the first three layers ($r = 3$) since the first three singular values of \mathbf{S} are much bigger than the rest. Only 5200 genes are selected overall. Among those, 3783, 2852, and 1187 genes are involved in the three layers, respectively. Heat maps of the three estimated layers are plotted in Figure 4. To better visualize the gene clustering, (1) all entries of the layers are divided by the maximum absolute value of the entries, (2) only the 5200 selected genes and the other 1000 randomly chosen unselected genes are plotted, (3) the genes in the figure are sorted hierarchically: firstly, the genes are sorted based on the ascending order of the entries of $\hat{\mathbf{u}}_1$, which automatically forms three gene groups according to the sign of the entries in $\hat{\mathbf{u}}_1$; secondly, within each group, we sort the genes bases on $\hat{\mathbf{u}}_2$, then nine gene groups are formed; finally, the sorting procedure is repeated based on $\hat{\mathbf{u}}_3$. The horizontal lines in each panel reveal the four cancer types of the subjects. The vertical lines in each panel reveal the 1000 unselected genes at the second column. It is clear that the proposed method is capable of simultaneously linking sets of genes to sets of subjects. Interestingly, the associations between gene groups and

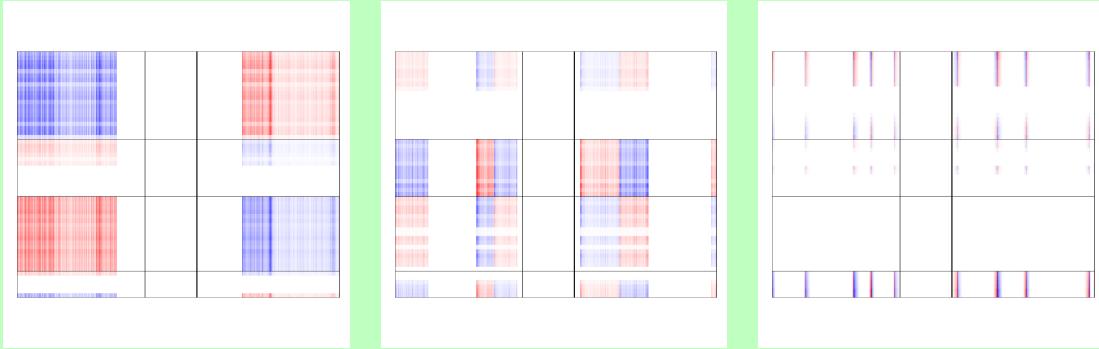


Figure 4: Estimated SVD layers by unsupervised biclustering. All entries of the layers are first divided by the maximum absolute value of the entries. Only the 5200 selected genes and the other 1000 randomly chosen unselected genes are plotted. The genes in the figure are sorted hierarchically: firstly, the genes are sorted based on the ascending order of the entries of $\hat{\mathbf{u}}_1$, which automatically forms three gene groups according to the sign of the entries in $\hat{\mathbf{u}}_1$; secondly, within each group, the genes are sorted based on $\hat{\mathbf{u}}_2$, then nine gene groups are formed; finally, the sorting procedure is repeated based on $\hat{\mathbf{u}}_3$. The horizontal lines in each panel reveal the four cancer types 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.

cancer types are clearly revealed in the estimated layers. For example, a very strong contrast between the Carcinoid group and the Normal group can be seen from the first layer, and another strong contrast between the Colon group and the Normal group can be seen from the second layer. A comparison between the original expression data \mathbf{S} to the sparse estimate $\hat{\mathbf{S}}$ (Figure 5) shows that our estimate $\hat{\mathbf{S}}$ successfully captures the basic structure of \mathbf{S} , and the zero-out areas in $\hat{\mathbf{S}}$ are indeed corresponding to most likely noninformative areas of \mathbf{S} . In this special case, our method actually shares very 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. Note that the number of genes in each layer we selected is slightly ~~bigger~~ than that of the SSVD, which may due to ~~the~~ difference in the form of penalty and regularization parameter selection. Further examination of the additionally selected genes shows that those genes form very similar clustering patterns as showed in Figure 4 and 5, although the signal is relatively weak. Therefore, we think those genes are all relevant and informative, which suggests that the SSVD method might have a larger false negative rate in this case.

The above ~~method~~ is an unsupervised learning tool, since the information of the available subject cancer type is not used. This may become a disadvantage when the primary interest is to identify gene-cancer type associations. In practice, expression heterogeneity can arise from various sources other than the cancer type factor, and these additional factors, which may be unknown or unmeasured, could cause within-group variations. In unsupervised learning, the subjects within a certain group are

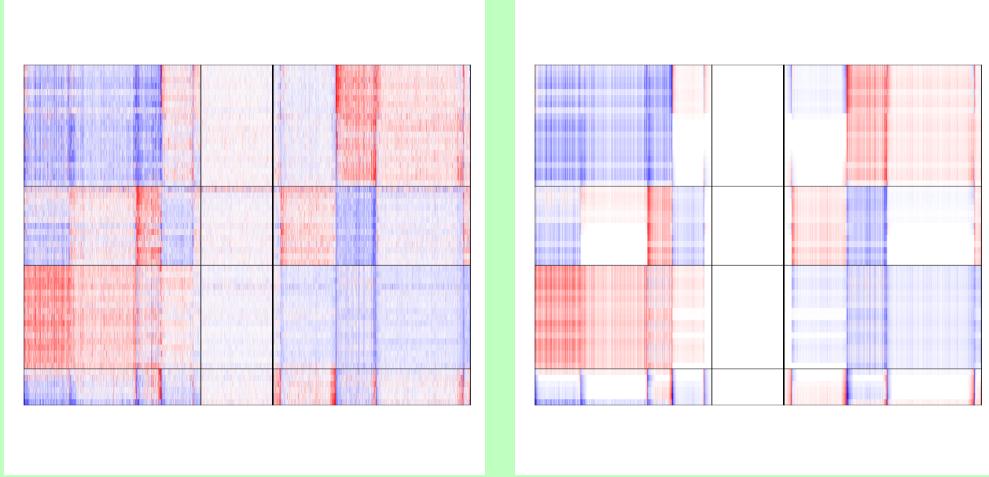


Figure 5: The original expression matrix (left) and the sparse estimate by unsupervised biclustering (right). All settings are the same as in Figure 4.

allowed to have differential responses to a set of genes. The consequence is that the estimated layers can be hard to interpret with respect to the group information, due to the failure of taking care of the within-group variations, which may in turn yield higher FDR or FNR. Figure 6, in which the estimated subject effects of the three SVD layers by the unsupervised biclustering are ~~sequential~~ shown, demonstrates such a problem in this lung cancer example. It shows that the within-group variations can be quite large, and sometimes may provide irrelevant or even ~~controversial~~ information about gene-cancer associations. Particularly, in the third SVD layer, some of the subjects of the Carcinoid group ~~have~~ positive responses, while ~~the~~ other in this group ~~do not~~ or ~~even~~ respond ~~conversely~~. Although such information may be valuable in that it suggests possible sub-grouping structure in the Carcinoid group, it is irrelevant on how to distinguish the four known cancer types.

In such situations, we believe that supervised learning may be preferable. The proposed sparse reduced-rank regression method can be easily turned to a supervised ~~learn~~ tool by incorporating cancer type information ~~about the subjects~~ to the covariate matrix \mathbf{G} . For this particular example, we could let \mathbf{G} be a 4×56 matrix such that

$$\mathbf{G} = \begin{bmatrix} \mathbf{1}_{20}^T & 0 & 0 & 0 \\ 0 & \mathbf{1}_{13}^T & 0 & 0 \\ 0 & 0 & \mathbf{1}_{17}^T & 0 \\ 0 & 0 & 0 & \mathbf{1}_6^T \end{bmatrix}.$$

In this setting, the coefficient matrix \mathbf{C} becomes a 12625×4 matrix, and each of its 1×4 right singular vectors can be regarded as representing group effects rather than individual subject effects, while each of its left singular vectors still represents the gene effects, with respect to a certain group or a ~~contrasts~~ of different groups. By doing so,

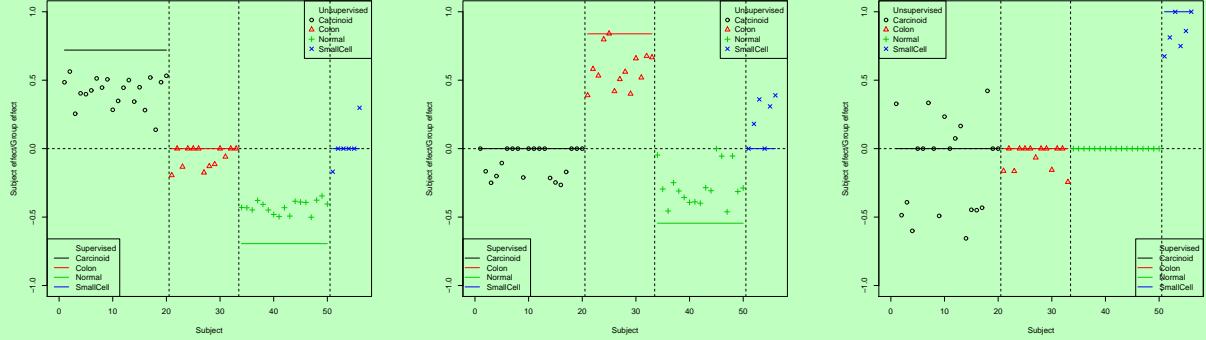


Figure 6: Estimated subject (group) effects by unsupervised (supervised) biclustering. Three SVD layers are shown sequentially.

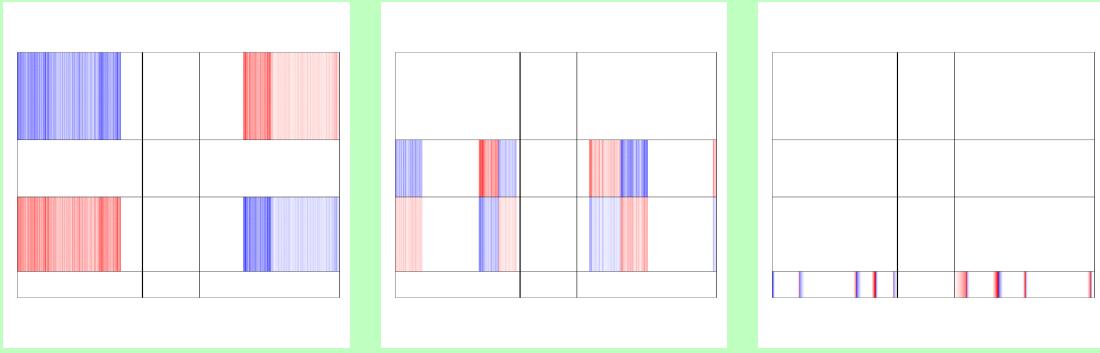


Figure 7: Estimated SVD layers by supervised biclustering. All settings are the same as in Figure 4.

the SVD is supervised such that it is forced to extract only the meaningful associations with respect to the cancer-type factor. Thus, this method is expected to be more robust than the unsupervised learning method. Since the \mathbf{G} matrix is still orthogonal, the computation stays to be very efficient. We then perform the supervised biclustering using the iterative exclusive-extraction method. Again, we consider only three layers since the first three singular values of $\tilde{\mathbf{C}}$ are ~~much bigger than the rest~~. Only 4663 genes are selected overall. Among those, 3507, 2231, and 1089 genes are involved in the three layers respectively. The estimated group effects of the three SVD layers are sequential, shown in 6. Heat maps of the three estimated layers are plotted in Figure 7, and a comparison between the original expression data \mathbf{S} to the estimate $\hat{\mathbf{S}}$ is shown in Figure 8. By supervised biclustering, more than one thousand genes are eliminated in the three layers  and only information about gene-cancer type associations are extracted and kept.

In gene expression studies, expression heterogeneity due to technical, genetic, envi-

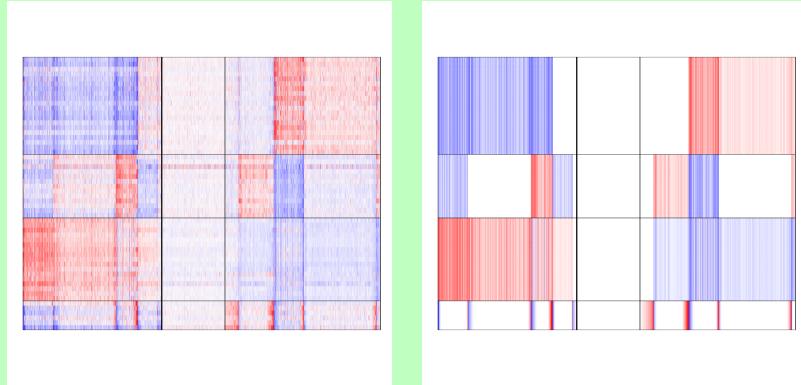


Figure 8: The original expression matrix (left) and the estimate by supervised biclustering (right). All settings are the same as in Figure 4.

ronmental, or demographic variables is very common. It is desirable to adjust for these covariate effects or “unwanted” variations while studying the clustering with respect to the primary variable, e.g. cancer type. Our methodology can be further extended for these needs. In general, we consider the reduced-rank regression model with two sets of regressors,

$$\mathbf{s}_t = \mathbf{C}\mathbf{g}_t + \mathbf{D}\mathbf{z}_t + \mathbf{e}_t, \quad t = 1, \dots, T, \quad (4.3)$$

where \mathbf{g}_t is constructed from the primary variable, \mathbf{z}_t is a $p \times 1$ vector of additional variables measured on the t th subject, \mathbf{D} is a $m \times p$ coefficient matrix that may be of full rank, and the other terms are defined as in model (1.1). This model was first suggested in the seminal work of Anderson (1951), and was studied by Reinsel and Velu (1998, Chapter 3) under classical least-square setting. Here under our regularized regression framework, the above extension adds no significant difficulty in estimation. One could still use a block coordinate decent algorithm to update \mathbf{C} and \mathbf{D} iteratively until converge.

5 Theoretical Property

We first recall or prove some useful results for the classical reduced-rank regression.

Proposition 1. (Reinsel and Velu, 1998): For the model (1.1), suppose $\text{rank}(\mathbf{C}) = r \leq \min(m, n)$. Then the minimizer of the objective (1.2) is given by

$$\tilde{\mathbf{C}} = \mathbf{A}\mathbf{A}^T\mathbf{S}\mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}$$

where $\mathbf{A} = [A_1, \dots, A_r]$ and A_j is the normalized eigenvector that corresponds to the j th largest eigenvalue of the matrix $\mathbf{S}\mathbf{G}^T(\mathbf{G}\mathbf{G}^T)^{-1}\mathbf{G}\mathbf{S}^T$ ($j = 1, \dots, r$). Moreover,

$$\sqrt{T}\text{vec}(\tilde{\mathbf{C}} - \mathbf{C}) \rightarrow_d N(0, \Sigma_c)$$

where the expression of Σ_c is given in (2.36) from Reinsel and Velu (1998).

Lemma 3. For the model (1.1), suppose some estimator of \mathbf{C} , say, $\tilde{\mathbf{C}}$, satisfies $\sqrt{T}\text{vec}(\tilde{\mathbf{C}} - \mathbf{C}) \rightarrow_d N(0, \Sigma_c)$. Let $\tilde{\mathbf{C}} = \sum_{k=1}^r \tilde{d}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^T$, $\mathbf{C} = \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^T$ be the SVD of $\tilde{\mathbf{C}}$ and \mathbf{C} respectively, where $d_1 > \dots > d_r > 0$. Then $\sqrt{T}(\tilde{d}_k - d_k)$, $\sqrt{T}(\tilde{\mathbf{u}}_k - \mathbf{u}_k)$ and $\sqrt{T}(\tilde{\mathbf{v}}_k - \mathbf{v}_k)$ for $k = 1, \dots, r$ are all asymptotically normally distributed with zero mean.

Proof: Here we only prove the asymptotical normality for $\tilde{\mathbf{u}}_k$, and the prove for \tilde{d}_k and $\tilde{\mathbf{v}}_k$ are similar. Recall that $\tilde{\mathbf{u}}_k$ (\tilde{d}_k) and \mathbf{u}_k (d_k) are the eigenvectors (eigenvalues) of $\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T$ and $\mathbf{C}\mathbf{C}^T$, respectively. With the use of perturbation expansion of matrices (Izenman, 1975), $\tilde{\mathbf{u}}_k$ can be expanded around \mathbf{u}_k to give

$$\begin{aligned} \sqrt{T}(\tilde{\mathbf{u}}_k - \mathbf{u}_k) &= \sqrt{T} \sum_{i \neq k}^r \frac{1}{d_k^2 - d_i^2} \mathbf{u}_i \mathbf{u}_k^T (\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T) \mathbf{u}_i \\ &\quad + o_p(\sqrt{T}\text{vec}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T)) \\ &= \sum_{i \neq k}^r \frac{1}{d_k^2 - d_i^2} (\mathbf{u}_i^T \otimes \mathbf{u}_i \mathbf{u}_k^T) \sqrt{T}\text{vec}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T) \\ &\quad + o_p(\sqrt{T}\text{vec}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T)). \end{aligned}$$

By the fact that $\sqrt{T}\text{vec}(\tilde{\mathbf{C}} - \mathbf{C}) \rightarrow_d N(0, \Sigma_c)$ and

$$\sqrt{T}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T) = \sqrt{T}(\tilde{\mathbf{C}} - \mathbf{C})\mathbf{C}^T + \sqrt{T}\mathbf{C}(\tilde{\mathbf{C}} - \mathbf{C})^T + \sqrt{T}(\tilde{\mathbf{C}} - \mathbf{C})(\tilde{\mathbf{C}} - \mathbf{C})^T,$$

We have

$$\sqrt{T}\text{vec}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T) = (\mathbf{C} \otimes \mathbf{I}_m) \sqrt{T}\text{vec}(\tilde{\mathbf{C}} - \mathbf{C}) + (\mathbf{I}_n \otimes \mathbf{C}) \sqrt{T}\text{vec}(\tilde{\mathbf{C}}^T - \mathbf{C}^T) + o_p(1).$$

Therefore $\sqrt{T}\text{vec}(\tilde{\mathbf{C}}\tilde{\mathbf{C}}^T - \mathbf{C}\mathbf{C}^T)$ is asymptotically normally distributed with zero mean. It then follows that $\sqrt{T}(\tilde{\mathbf{u}}_k - \mathbf{u}_k)$ is also asymptotically normally distributed with zero

mean.

Remark: In Lemma 3, for simplicity we do not give explicit expression of the asymptotical covariance matrix. However, it can be easily obtained by Delta method.

In the following, we first investigate the theoretical property of the proposed regularized estimators for the unit-rank case. Then we extend the theory to the general case. We consider the following conditions:

C1. $\frac{1}{T}\mathbf{G}\mathbf{G}^T \rightarrow \mathbf{K}$ where \mathbf{K} is a positive definite matrix. Let

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix},$$

where \mathbf{K}_{11} is a $n_0 \times n_0$ matrix.

C2. The $m \times 1$ vector of random errors \mathbf{e}_t is independently and identically distributed (i.i.d) with mean vector $E(\mathbf{e}_t) = 0$ and covariance matrix $Cov(\mathbf{e}_t) = \Sigma_e$, an $m \times m$ positive-definite matrix. Let

$$\Sigma_e = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

where Σ_{11} is a $m_0 \times m_0$ matrix.

C3. $\frac{\lambda_T}{\sqrt{T}} \rightarrow 0$, $\frac{\lambda_T}{\sqrt{T}} T^{\frac{\gamma}{2}} \rightarrow \infty$ with $\gamma > 0$.

5.1 Unit-rank Case

Suppose the true model is given as (1.1), where $\mathbf{C} \in \Omega = \bigcup_{i=1}^m \Omega_i$, and

$$\Omega_i = \{\mathbf{u}\mathbf{v}^T; \mathbf{u} \in R^m \text{ with } u_i = 1, \mathbf{v} \in R^n \text{ and } \mathbf{v} \neq 0\}.$$

Here for simplicity the singular value is absorbed into the singular vectors. Each Ω_i is composed of nonzero rank-1 matrices whose i th entry of its left singular vector is nonzero. Therefore the matrix space Ω consists of all the nonzero rank-1 matrix. Without loss of generality, in the following we assume $\mathbf{C} \in \Omega_1$ with $\mathbf{C} = \mathbf{u}^* \mathbf{v}^{*T}$ and $u_1^* = 1$. Let $\mathcal{A} = \{i : u_i^* \neq 0\}$ and $\mathcal{B} = \{j : v_j^* \neq 0\}$. Without loss of generality, we assume $\mathcal{A} = \{1, \dots, m_0\}$ and $\mathcal{B} = \{1, \dots, n_0\}$ where $m_0 \leq m$ and $n_0 \leq n$. Denote $\mathcal{A} - x = \{i; i \in \mathcal{A}, i \neq x\}$ and $\mathcal{A}\mathcal{B} = \{(i, j); i \in \mathcal{A}, j \in \mathcal{B}\}$.

Let $Q(\mathbf{u}, \mathbf{v})$ denote the objective function as (2.1), i.e.

$$Q_T(\mathbf{u}, \mathbf{v}) = \text{tr}[(\mathbf{S} - \mathbf{u}\mathbf{v}^T \mathbf{G})(\mathbf{S} - \mathbf{u}\mathbf{v}^T \mathbf{G})^T] + \lambda_T \sum_{i=1}^m \sum_{j=1}^n w_{ij} |u_i v_j| \quad (5.1)$$

and let $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)}) = \arg\min Q_T(\mathbf{u}, \mathbf{v})$. Theorem 1-3 study the asymptotic properties of $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ as the sample size $T \rightarrow \infty$.

Theorem 1. Suppose condition **C1** and **C2** are satisfied, and suppose $\frac{\lambda_T}{\sqrt{T}} \rightarrow \lambda_0 \geq 0$ as $T \rightarrow \infty$. Then there exists a local minimizer $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ of $Q_T(\mathbf{u}, \mathbf{v})$ in (5.1) such that $\|\hat{\mathbf{u}}^{(T)} - \mathbf{u}^*\| = O_p(T^{-\frac{1}{2}})$ and $\|\hat{\mathbf{v}}^{(T)} - \mathbf{v}^*\| = O_p(T^{-\frac{1}{2}})$.

Proof: We consider a neighborhood of \mathbf{C} in Ω_1 of radius $r > 0$:

$$\mathcal{N}(\mathbf{C}, r) = \{(\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a}, \mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b}), \mathbf{a} \in R^m \text{ with } a_1 = 0; \mathbf{b} \in R^n; \|\mathbf{a}\| \leq r; \|\mathbf{b}\| \leq r\}.$$

In the following, we let $\mathbf{a} \in R^m$ with $a_1 = 0$ and $\mathbf{b} \in R^n$ unless otherwise noted. For any $(\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a}, \mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b}) \in \mathcal{N}(\mathbf{C}, r)$, we have

$$\begin{aligned} & Q_T(\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a}, \mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b}) \\ &= \text{tr}[(\mathbf{S} - (\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a})(\mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b})^T \mathbf{G})(\mathbf{S} - (\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a})(\mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b})^T \mathbf{G})^T] \\ &+ \lambda_T \sum_{i=1}^m \sum_{j=1}^n w_{ij} |u_i^* + \frac{1}{\sqrt{T}}a_i| |v_j^* + \frac{1}{\sqrt{T}}b_j|. \end{aligned}$$

We want to show that for any given $\epsilon > 0$, there exists a large enough constant r such that

$$P\{\inf_{\|\mathbf{a}\|=\|\mathbf{b}\|=r} Q_T(\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a}, \mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b}) > Q_T(\mathbf{u}^*, \mathbf{v}^*)\} \geq 1 - \epsilon.$$

This implies that with probability at least $1 - \epsilon$ there exists a local minimum in the interior of the ball $\mathcal{N}(\mathbf{C}, r)$. Hence, there exists a local minimizer such that $\|\hat{\mathbf{u}} - \mathbf{u}^*\| = O_p(T^{-1/2})$ and $\|\hat{\mathbf{v}} - \mathbf{v}^*\| = O_p(T^{-1/2})$.

Define $\Psi_T(\mathbf{a}, \mathbf{b}) \equiv Q_T(\mathbf{u}^* + \frac{1}{\sqrt{T}}\mathbf{a}, \mathbf{v}^* + \frac{1}{\sqrt{T}}\mathbf{b}) - Q_T(\mathbf{u}^*, \mathbf{v}^*)$. We have

$$\begin{aligned} \Psi_T(\mathbf{a}, \mathbf{b}) &= -2\text{tr}[(\mathbf{u}^*\mathbf{b}^T + \mathbf{a}\mathbf{v}^T + \frac{1}{\sqrt{T}}\mathbf{a}\mathbf{b}^T) \frac{1}{\sqrt{T}} \mathbf{G} \mathbf{E}^T] \\ &+ \text{tr}[(\mathbf{u}^*\mathbf{b}^T + \mathbf{a}\mathbf{v}^T + \frac{1}{\sqrt{T}}\mathbf{a}\mathbf{b}^T) \frac{\mathbf{G}\mathbf{G}^T}{T} (\mathbf{u}^*\mathbf{b}^T + \mathbf{a}\mathbf{v}^T + \frac{1}{\sqrt{T}}\mathbf{a}\mathbf{b}^T)^T] \\ &+ \frac{\lambda_T}{\sqrt{T}} \sum_{i=1}^m \sum_{j=1}^n w_{ij} \sqrt{T} [|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_i v_j^* + \frac{1}{T}a_i b_j| - |u_i^*v_j^*|]. \end{aligned}$$

By some algebra, we have

$$\begin{aligned}
\Psi_T(\mathbf{a}, \mathbf{b}) = & -2\text{vec}^T(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T + \frac{1}{\sqrt{T}}\mathbf{b}\mathbf{a}^T)\text{vec}(\frac{1}{\sqrt{T}}\mathbf{G}\mathbf{E}^T) \\
& + \text{vec}^T(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T + \frac{1}{\sqrt{T}}\mathbf{b}\mathbf{a}^T)(\mathbf{I}_m \otimes \frac{\mathbf{G}\mathbf{G}^T}{T})\text{vec}(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T + \frac{1}{\sqrt{T}}\mathbf{b}\mathbf{a}^T) \\
& + \frac{\lambda_T}{\sqrt{T}} \sum_{i=1}^m \sum_{j=1}^n w_{ij} \sqrt{T} [|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_i^*v_j^* + \frac{1}{T}a_i^*b_j| - |u_i^*v_j^*|]. \quad (5.2)
\end{aligned}$$

We know $\text{vec}(\frac{1}{\sqrt{T}}\mathbf{G}\mathbf{E}^T) \rightarrow_d N(0, \Sigma_e \otimes \mathbf{K})$ and

$$\sqrt{T} [|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_i^*v_j^* + \frac{1}{T}a_i^*b_j| - |u_i^*v_j^*|] \rightarrow \begin{cases} \text{sign}(u_i^*v_j^*)(u_i^*b_j + a_i^*v_j^*) & u_i^*v_j^* \neq 0 \\ |a_i^*v_j^*| & u_i^* = 0, v_j^* \neq 0 \\ |u_i^*b_j| & u_i^* \neq 0, v_j^* = 0 \\ 0 & u_i^* = 0, v_j^* = 0 \end{cases}$$

as $T \rightarrow \infty$. Therefore when T is sufficiently large,

$$\begin{aligned}
\Psi_T(\mathbf{a}, \mathbf{b}) \geq & -2\text{vec}^T(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T)\text{vec}(\frac{1}{\sqrt{T}}\mathbf{G}\mathbf{E}^T) \\
& + \text{vec}^T(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T)(\mathbf{I}_m \otimes \frac{\mathbf{G}\mathbf{G}^T}{T})\text{vec}(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T) \\
& + \frac{\lambda_T}{\sqrt{T}} \sum_{i=1}^m \sum_{j=1}^n w_{ij} \text{sign}(u_i^*v_j^*)(u_i^*b_j + a_i^*v_j^*)
\end{aligned}$$

Now it suffices to show for a sufficiently large r , the second term on the right-hand side dominates both the first and the third term in $\|\mathbf{a}\| = \|\mathbf{b}\| = r$. Given the facts that $\text{vec}(\frac{1}{\sqrt{T}}\mathbf{G}\mathbf{E}^T) \rightarrow_d N(0, \Sigma_e \otimes \mathbf{K})$, $\mathbf{I}_m \otimes \frac{\mathbf{G}\mathbf{G}^T}{T} \rightarrow \mathbf{I}_m \otimes \mathbf{K}$, $\frac{\lambda_T}{\sqrt{T}} \rightarrow \lambda_0$ and $w_{ij} \rightarrow_p |u_i^*v_j^*|^{-\gamma}$ as $T \rightarrow \infty$, it then suffices to show that for a sufficiently large r , denoted as r_T^* , $\|\text{vec}(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T)\|^2$ dominates $\|\text{vec}(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T)\|$. Since $a_1 = 0$ and $u_1^* = 1$, the first column of $\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T$ is \mathbf{b} . It then follows that

$$\|\text{vec}(\mathbf{b}\mathbf{u}^{*T} + \mathbf{v}^*\mathbf{a}^T)\|^2 = r^2(1 + f(\frac{\mathbf{a}}{r}, \frac{\mathbf{b}}{r})),$$

where $f(\cdot, \cdot)$ is a bounded and continuous function on the unit-sphere

$$\{(\mathbf{a}, \mathbf{b}); \mathbf{a} \in R^m \text{ with } a_1 = 0 \text{ and } \|\mathbf{a}\| = 1, \mathbf{b} \in R^n \text{ with } \|\mathbf{b}\| = 1\}.$$

This guarantees the existence of r_T^* and hence completes the proof of the theorem.

Remark: By the law of iterated logarithm, $\frac{1}{\sqrt{T}}\mathbf{G}\mathbf{E}^T \leq K\sqrt{\log \log T}$ a.s. for some $K > 0$. It then follows that the radius r_T^* of the neighborhood $\mathcal{N}(\mathbf{C}, r_T^*)$, in which the local minimum is guaranteed to exist, is of the order $O(\sqrt{\log \log T})$. Therefore as

$T \rightarrow \infty$, the neighborhood $\mathcal{N}(\mathbf{C}, r_T^*)$ expands to the whole parameter space Ω_T .

Theorem 2. Suppose condition **C1-C3** are satisfied. Let $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ be the local minimizer of $Q_T(\mathbf{u}, \mathbf{v})$ in (5.1) as found in Theorem 1. Then $\sqrt{T}(\hat{\mathbf{u}}_{\mathcal{A}-1}^{(T)} - \mathbf{u}_{\mathcal{A}-1}^*)$ and $\sqrt{T}(\hat{\mathbf{v}}_{\mathcal{B}}^{(T)} - \mathbf{v}_{\mathcal{B}}^*)$ are both asymptotically normally distributed.

Proof: Again, we consider $\Psi_T(\mathbf{a}, \mathbf{b})$ in (5.2). We know that $\text{vec}(\frac{1}{\sqrt{T}}\mathbf{G}\mathbf{E}^T) \rightarrow_d N(0, \Sigma_e \otimes \mathbf{K})$ and $\mathbf{I}_m \otimes \frac{\mathbf{G}\mathbf{G}^T}{T} \rightarrow \mathbf{I}_m \otimes \mathbf{K}$. Now consider the third term:

- If $u_i^*v_j^* \neq 0$:

$$\begin{aligned} w_{ij} &\rightarrow_p |u_i^*v_j^*|^{-\gamma}; \\ \sqrt{T}[|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_iv_j^* + \frac{1}{T}a_ib_j| - |u_i^*v_j^*|] &\rightarrow \text{sign}(u_i^*v_j^*)(u_i^*b_j + a_iv_j^*); \\ \frac{\lambda_T}{\sqrt{T}} &\rightarrow 0; \\ \Rightarrow \frac{\lambda_T}{\sqrt{T}}w_{ij}\sqrt{T}[|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_iv_j^* + \frac{1}{T}a_ib_j| - |u_i^*v_j^*|] &\rightarrow_p 0. \end{aligned}$$

- If $u_i^* = 0, v_j^* \neq 0$:

$$\begin{aligned} \frac{\lambda_T}{\sqrt{T}}w_{ij} &= \frac{\lambda_T}{\sqrt{T}}T^{\frac{\gamma}{2}}|\sqrt{T}\tilde{c}_{ij}|^{-\gamma} \rightarrow_p \infty; \\ \sqrt{T}[|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_iv_j^* + \frac{1}{T}a_ib_j| - |u_i^*v_j^*|] &\rightarrow |a_iv_j^*|; \\ \Rightarrow \frac{\lambda_T}{\sqrt{T}}w_{ij}\sqrt{T}[|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_iv_j^* + \frac{1}{T}a_ib_j| - |u_i^*v_j^*|] &\rightarrow_p \infty \text{ if } a_i \neq 0. \end{aligned}$$

- If $u_i^* \neq 0, v_j^* = 0$: by similar argument,

$$\frac{\lambda_T}{\sqrt{T}}w_{ij}\sqrt{T}[|u_i^*v_j^* + \frac{1}{\sqrt{T}}u_i^*b_j + \frac{1}{\sqrt{T}}a_iv_j^* + \frac{1}{T}a_ib_j| - |u_i^*v_j^*|] \rightarrow_p \infty \text{ if } b_i \neq 0.$$

Therefore,

$$\Psi_T(\mathbf{a}, \mathbf{b}) \rightarrow_d \Psi(\mathbf{a}, \mathbf{b}) = \begin{cases} -2\mathbf{z}^T \mathbf{w}_{\mathcal{A}\mathcal{B}} + \mathbf{z}^T (\mathbf{I}_{m_0} \otimes \mathbf{K}_{11}) \mathbf{z} & a_i = 0, i \notin \mathcal{A}; b_j = 0, j \notin \mathcal{B}; \\ \infty & \text{otherwise.} \end{cases}$$

where $\mathbf{z} = \text{vec}(\mathbf{b}_{\mathcal{B}}\mathbf{u}_{\mathcal{A}}^{*T} + \mathbf{v}_{\mathcal{B}}^*\mathbf{a}_{\mathcal{A}}^T)$ and $\mathbf{w}_{\mathcal{A}\mathcal{B}} \sim N(0, \Sigma_{11} \otimes \mathbf{K}_{11})$.

Next we show that $\Psi(\mathbf{a}, \mathbf{b})$ has a unique minimum which is denoted as $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$. Obviously, $\forall i \notin \mathcal{A}, \hat{a}_i = 0$, and $\forall j \notin \mathcal{B}, \hat{b}_j = 0$. Now consider $\Psi_{\mathcal{A}\mathcal{B}}(\mathbf{a}_{\mathcal{A}}, \mathbf{b}_{\mathcal{B}}) \equiv -2\mathbf{z}^T \mathbf{w}_{\mathcal{A}\mathcal{B}} +$

$\mathbf{z}^T(\mathbf{I}_{m_0} \otimes \mathbf{K}_{11})\mathbf{z}$. We know $\text{vec}(\mathbf{b}_\mathcal{B}\mathbf{u}_\mathcal{A}^{*T}) = (\mathbf{u}_\mathcal{A}^* \otimes \mathbf{I}_{n_0})\mathbf{b}_\mathcal{B}$, and $\text{vec}(\mathbf{v}_\mathcal{B}^*\mathbf{a}_\mathcal{A}^T) = (\mathbf{I}_{m_0} \otimes \mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}$, then

$$\begin{aligned}\Psi_{\mathcal{A}\mathcal{B}}(\mathbf{a}_\mathcal{A}, \mathbf{b}_\mathcal{B}) &= -2\mathbf{b}_\mathcal{B}^T(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{I}_{n_0})\mathbf{w}_{\mathcal{A}\mathcal{B}} + (\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*)\mathbf{b}_\mathcal{B}^T\mathbf{K}_{11}\mathbf{b}_\mathcal{B} \\ &\quad + 2\mathbf{b}_\mathcal{B}^T(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A} \\ &\quad - 2\mathbf{a}_\mathcal{A}^T(\mathbf{I}_{m_0} \otimes \mathbf{v}_\mathcal{B}^{*T})\mathbf{w}_{\mathcal{A}\mathcal{B}} + (\mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}^T\mathbf{a}_\mathcal{A}.\end{aligned}$$

To find the unique minimum, we first assume $\mathbf{a}_\mathcal{A}$ is known. Then

$$\begin{aligned}\Psi_{\mathcal{A}\mathcal{B}}(\mathbf{a}_\mathcal{A}, \mathbf{b}_\mathcal{B} | \mathbf{a}_\mathcal{A}) &= -2\mathbf{b}_\mathcal{B}^T[(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{I}_{n_0})\mathbf{w}_{\mathcal{A}\mathcal{B}} - (\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}] \\ &\quad + (\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*)\mathbf{b}_\mathcal{B}^T\mathbf{K}_{11}\mathbf{b}_\mathcal{B} + \text{const},\end{aligned}$$

which is a convex function of $\mathbf{b}_\mathcal{B}$ because \mathbf{K}_{11} is positive definite. The unique minimizer is given by

$$\hat{\mathbf{b}}_\mathcal{B} = \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}\mathbf{K}_{11}^{-1}[(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{I}_{n_0})\mathbf{w}_{\mathcal{A}\mathcal{B}} - (\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}]$$

We plug in this expression to the original objective function $\Psi_{\mathcal{A}\mathcal{B}}(\mathbf{a}_\mathcal{A}, \mathbf{b}_\mathcal{B})$ then

$$\begin{aligned}\Psi_{\mathcal{A}\mathcal{B}}(\mathbf{a}_\mathcal{A}, \mathbf{b}_\mathcal{B} | \mathbf{b}_\mathcal{B} = \hat{\mathbf{b}}_\mathcal{B}) &= -2\mathbf{a}_\mathcal{A}^T(\mathbf{I}_{m_0} \otimes \mathbf{v}_\mathcal{B}^{*T})\mathbf{w}_{\mathcal{A}\mathcal{B}} + (\mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}^T\mathbf{a}_\mathcal{A} \\ &\quad - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}[\mathbf{w}_{\mathcal{A}\mathcal{B}}^T(\mathbf{u}_\mathcal{A}^* \otimes \mathbf{I}_{n_0}) - \mathbf{a}_\mathcal{A}^T(\mathbf{u}_\mathcal{A}^* \otimes \mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}^T)]\mathbf{K}_{11}^{-1}[(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{I}_{n_0})\mathbf{w}_{\mathcal{A}\mathcal{B}} - (\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}] \\ &= -2\mathbf{a}_\mathcal{A}^T[\mathbf{I}_{m_0} \otimes \mathbf{v}_\mathcal{B}^{*T} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}(\mathbf{u}_\mathcal{A}^* \otimes \mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}^T)\mathbf{K}_{11}^{-1}(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{I}_{n_0})]\mathbf{w}_{\mathcal{A}\mathcal{B}} \\ &\quad + \mathbf{a}_\mathcal{A}^T[(\mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{I}_{m_0} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}(\mathbf{u}_\mathcal{A}^* \otimes \mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}^T)\mathbf{K}_{11}^{-1}(\mathbf{u}_\mathcal{A}^{*T} \otimes \mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)]\mathbf{a}_\mathcal{A} + \text{const} \\ &= -2\mathbf{a}_\mathcal{A}^T[(\mathbf{I}_{m_0} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}\mathbf{u}_\mathcal{A}^*\mathbf{u}_\mathcal{A}^{*T}) \otimes \mathbf{v}_\mathcal{B}^{*T}]\mathbf{w}_{\mathcal{A}\mathcal{B}} \\ &\quad + (\mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_\mathcal{A}^T(\mathbf{I}_{m_0} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}\mathbf{u}_\mathcal{A}^*\mathbf{u}_\mathcal{A}^{*T})\mathbf{a}_\mathcal{A} + \text{const}\end{aligned}$$

Recall $a_1 = 0$ and $u_1^* = 1$, then

$$\begin{aligned}\Psi_{\mathcal{A}\mathcal{B}}(\mathbf{a}_\mathcal{A}, \mathbf{b}_\mathcal{B} | \mathbf{b}_\mathcal{B} = \hat{\mathbf{b}}_\mathcal{B}) &= -2\mathbf{a}_{\mathcal{A}-1}^T[(\mathbf{I}_{m_0-1} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}\mathbf{u}_{\mathcal{A}-1}^*\mathbf{u}_{\mathcal{A}-1}^{*T}) \otimes \mathbf{v}_\mathcal{B}^{*T}]\mathbf{w}_{(\mathcal{A}-1)\mathcal{B}} \\ &\quad + (\mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*)\mathbf{a}_{\mathcal{A}-1}^T(\mathbf{I}_{m_0-1} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}\mathbf{u}_{\mathcal{A}-1}^*\mathbf{u}_{\mathcal{A}-1}^{*T})\mathbf{a}_{\mathcal{A}-1} + \text{const}\end{aligned}$$

which is a convex function of $\mathbf{a}_{\mathcal{A}-1}$ because $(\mathbf{I}_{m_0-1} - \frac{1}{\mathbf{u}_\mathcal{A}^{*T}\mathbf{u}_\mathcal{A}^*}\mathbf{u}_{\mathcal{A}-1}^*\mathbf{u}_{\mathcal{A}-1}^{*T})$ is positive definite. The unique minimizer is given by

$$\hat{\mathbf{a}}_{\mathcal{A}-1} = \frac{1}{\mathbf{v}_\mathcal{B}^{*T}\mathbf{K}_{11}\mathbf{v}_\mathcal{B}^*}(\mathbf{I}_{m_0-1} \otimes \mathbf{v}_\mathcal{B}^{*T})\mathbf{w}_{(\mathcal{A}-1)\mathcal{B}}.$$

Then it follows that

$$\hat{\mathbf{b}}_{\mathcal{B}} = \frac{1}{\mathbf{u}_{\mathcal{A}}^{*T} \mathbf{u}_{\mathcal{A}}^*} [\mathbf{u}_{\mathcal{A}}^{*T} \otimes \mathbf{K}_{11}^{-1} - \frac{1}{\mathbf{v}_{\mathcal{B}}^{*T} \mathbf{K}_{11} \mathbf{v}_{\mathcal{B}}^*} (\mathbf{u}_{\mathcal{A}}^{*T} \mathbf{J} \otimes \mathbf{v}_{\mathcal{B}}^* \mathbf{v}_{\mathcal{B}}^{*T})] \mathbf{w}_{\mathcal{A}\mathcal{B}}$$

where

$$\mathbf{J} = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{m_0-1} \end{bmatrix}.$$

Finally, we show the asymptotic normality results. Let $\hat{\mathbf{a}}^{(T)} = \sqrt{T}(\hat{\mathbf{u}}^{(T)} - \mathbf{u}^*)$ and $\hat{\mathbf{b}}^{(T)} = \sqrt{T}(\hat{\mathbf{v}}^{(T)} - \mathbf{v}^*)$. Let $\mathcal{H} = \{(a, b); a \in R^m \text{ with } a_1 = 0, b \in R^n\}$ and

$$\mathcal{H}_T = \{(a, b); a \in R^m \text{ with } a_1 = 0, b \in R^n, \|a\| \leq r_T^*, \|b\| \leq r_T^*\}.$$

where the radius r_T^* is defined as in the proof of Theorem 2. We have

- $\Psi_T \rightarrow_d \Psi$ for any compact set of \mathcal{H} ;
- The limit process Ψ has continuous sample path and ~~unique point of minima~~ $(\hat{\mathbf{a}}, \hat{\mathbf{b}})$;
- $\mathcal{H}_T \rightarrow \mathcal{H}$ as $T \rightarrow \infty$ since $r_T^* = O(\sqrt{\log \log T})$, and $\Psi_T(\hat{\mathbf{a}}^{(T)}, \hat{\mathbf{b}}^{(T)}) \leq \Psi_T(\mathcal{H}_T) + o_p(1)$;
- The sequence $(\hat{\mathbf{a}}^{(T)}, \hat{\mathbf{b}}^{(T)})$ is uniformly ~~tight~~ \square .

Therefore, by the argmax theorem (van der Vaart, 2000), we have $(\hat{\mathbf{a}}^{(T)}, \hat{\mathbf{b}}^{(T)}) \rightarrow_d (\hat{\mathbf{a}}, \hat{\mathbf{b}})$, i.e.

$$\begin{aligned} \sqrt{T}(\hat{\mathbf{u}}_{\mathcal{A}-1}^{(T)} - \mathbf{u}_{\mathcal{A}-1}^*) &\rightarrow_d \frac{1}{\mathbf{v}_{\mathcal{B}}^{*T} \mathbf{K}_{11} \mathbf{v}_{\mathcal{B}}^*} (\mathbf{I}_{m_0-1} \otimes \mathbf{v}_{\mathcal{B}}^{*T}) \mathbf{w}_{(\mathcal{A}-1)\mathcal{B}}; \\ \sqrt{T}(\hat{\mathbf{v}}_{\mathcal{B}}^{(T)} - \mathbf{v}_{\mathcal{B}}^*) &\rightarrow_d \frac{1}{\mathbf{u}_{\mathcal{A}}^{*T} \mathbf{u}_{\mathcal{A}}^*} [\mathbf{u}_{\mathcal{A}}^{*T} \otimes \mathbf{K}_{11}^{-1} - \frac{1}{\mathbf{v}_{\mathcal{B}}^{*T} \mathbf{K}_{11} \mathbf{v}_{\mathcal{B}}^*} (\mathbf{u}_{\mathcal{A}}^{*T} \mathbf{J} \otimes \mathbf{v}_{\mathcal{B}}^* \mathbf{v}_{\mathcal{B}}^{*T})] \mathbf{w}_{\mathcal{A}\mathcal{B}}; \\ \sqrt{T}(\hat{\mathbf{u}}_{\mathcal{A}^c}^{(T)} - \mathbf{u}_{\mathcal{A}^c}^*) &\rightarrow_d 0; \\ \sqrt{T}(\hat{\mathbf{v}}_{\mathcal{B}^c}^{(T)} - \mathbf{v}_{\mathcal{B}^c}^*) &\rightarrow_d 0. \end{aligned}$$

The proof is completed. \square

Theorem 3. Suppose condition ~~C1-C~~ are satisfied. Let $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ be the local minimizer of $Q_T(\mathbf{u}, \mathbf{v})$ in (5.1) as found in Theorem 1. Let $\mathcal{A}_T = \{i : \hat{u}_i^{(T)} \neq 0\}$ and $\mathcal{B}_T = \{j : \hat{v}_j^{(T)} \neq 0\}$. Then $P(\mathcal{A}_T = \mathcal{A}) \rightarrow 1$ and $P(\mathcal{B}_T = \mathcal{B}) \rightarrow 1$ as $T \rightarrow \infty$.

Proof: According to the asymptotic normality result $\hat{\mathbf{u}}_{\mathcal{A}}^{(T)} \rightarrow_p \mathbf{u}_{\mathcal{A}}^*$ and $\hat{\mathbf{v}}_{\mathcal{B}}^{(T)} \rightarrow_p \mathbf{v}_{\mathcal{B}}^*$; thus $\forall i \in \mathcal{A}$, $P(i \in \mathcal{A}_T) \rightarrow 1$, and $\forall j \in \mathcal{B}$, $P(j \in \mathcal{B}_T) \rightarrow 1$. Then it suffices to show that

$\forall i \notin \mathcal{A}$, $P(i \in \mathcal{A}_T) \rightarrow 0$, and $\forall j \notin \mathcal{B}$, $P(j \in \mathcal{B}_T) \rightarrow 0$. In the following, for simplicity we write $\hat{\mathbf{u}}^{(T)} = \hat{\mathbf{u}}$ and $\hat{\mathbf{v}}^{(T)} = \hat{\mathbf{v}}$.

$\forall i \notin \mathcal{A}$, consider the event $i \in \mathcal{A}_T$. By the KKT optimality conditions, we know that

$$\frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{v}),i}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{v})} \hat{\mathbf{u}}) = \frac{1}{\sqrt{T}} \lambda_{(\mathbf{v})} w_{1,i} \quad (5.4)$$

where $\mathbf{X}_{(\mathbf{v})} = (\mathbf{X}_{(\mathbf{v}),1}, \dots, \mathbf{X}_{(\mathbf{v}),m}) = \mathbf{I}_m \otimes \mathbf{G}^T \hat{\mathbf{v}}$, $\lambda_{(\mathbf{v})} = \lambda_T \sum_{j=1}^n w_{2,j} |\hat{v}_j|$. Consider the left-hand side:

$$\begin{aligned} LHS &= \frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{v}),i}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{v})} \hat{\mathbf{u}}) \\ &= \frac{1}{\sqrt{T}} [\hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \mathbf{v}^* u_i^* - \hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \hat{\mathbf{v}} \hat{u}_i + \hat{\mathbf{v}}^T \mathbf{G} \mathbf{e}_{(i)}] \\ &= -\frac{\hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \hat{\mathbf{v}}}{T} \sqrt{T} \hat{u}_i + \frac{\hat{\mathbf{v}}^T \mathbf{G} \mathbf{e}_{(i)}}{\sqrt{T}} \\ &= O_p(1). \end{aligned}$$

Consider the right-hand side:

$$\begin{aligned} RHS &= \frac{1}{\sqrt{T}} \lambda_{(\mathbf{v})} w_{1,i} \\ &= \frac{\lambda_T}{\sqrt{T}} |\hat{\mathbf{v}}|^T \mathbf{w}_2 w_{1,i} \\ &= \frac{\lambda_T}{\sqrt{T}} T^{\frac{\gamma}{2}} |\hat{\mathbf{v}}|^T \mathbf{w}_2 \frac{1}{|\sqrt{T} \hat{u}_i|^\gamma} \\ &\rightarrow_p \infty. \end{aligned}$$

Therefore,

$$P(i \in \mathcal{A}_T) \leq P\left(\frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{v}),i}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{v})} \hat{\mathbf{u}}) = \frac{1}{\sqrt{T}} \lambda_{(\mathbf{v})} w_{1,i}\right) \rightarrow 0.$$

$\forall j \notin \mathcal{B}$, consider the event $j \in \mathcal{B}_T$. By the KKT optimality conditions, we know that

$$\frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{u}),j}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{u})} \hat{\mathbf{v}}) = \frac{1}{\sqrt{T}} \lambda_{(\mathbf{u})} w_{2,j} \quad (5.5)$$

where $\mathbf{X}_{(\mathbf{u})} = (\mathbf{X}_{(\mathbf{u}),1}, \dots, \mathbf{X}_{(\mathbf{u}),n}) = \hat{\mathbf{u}} \otimes \mathbf{G}^T$, $\lambda_{(\mathbf{u})} = \lambda_T \sum_{i=1}^m w_{1,i} |\hat{u}_i|$. Consider the left-

hand side:

$$\begin{aligned}
LHS &= \frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{u}),j}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{u})} \hat{\mathbf{v}}) \\
&= \frac{1}{\sqrt{T}} [\hat{\mathbf{u}}^T \mathbf{u}^* \mathbf{v}^{*T} \mathbf{G} \mathbf{g}_{(i)} - \hat{\mathbf{u}}^T \hat{\mathbf{u}} \hat{\mathbf{v}}^T \mathbf{G} \mathbf{g}_{(i)} + \hat{\mathbf{u}}^T \mathbf{E} \mathbf{g}_{(i)}] \\
&= \sqrt{T} \hat{\mathbf{u}}^T \mathbf{u}^* \mathbf{v}^{*T} \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} - \sqrt{T} \hat{\mathbf{u}}^T \hat{\mathbf{u}} \hat{\mathbf{v}}^T \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} + \frac{\hat{\mathbf{u}}^T \mathbf{E} \mathbf{g}_{(i)}}{\sqrt{T}} \\
&= \sqrt{T} (\hat{\mathbf{u}} - \mathbf{u}^*)^T \mathbf{u}^* \mathbf{v}^{*T} \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} - \hat{\mathbf{u}}^T \hat{\mathbf{u}} \sqrt{T} (\hat{\mathbf{v}}^T - \mathbf{v}^{*T}) \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} \\
&\quad - \sqrt{T} (\hat{\mathbf{u}}^T \hat{\mathbf{u}} - \mathbf{u}^{*T} \mathbf{u}^*) \mathbf{v}^{*T} \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} + \frac{\hat{\mathbf{u}}^T \mathbf{E} \mathbf{g}_{(i)}}{\sqrt{T}} \\
&= O_p(1).
\end{aligned}$$

Consider the right-hand side:

$$\begin{aligned}
RHS &= \frac{1}{\sqrt{T}} \lambda_{(\mathbf{u})} w_{2,j} \\
&= \frac{\lambda_T}{\sqrt{T}} |\hat{\mathbf{u}}|^T \mathbf{w}_1 w_{2,j} \\
&= \frac{\lambda_T}{\sqrt{T}} T^{\frac{\gamma}{2}} |\hat{\mathbf{u}}|^T \mathbf{w}_1 \frac{1}{|\sqrt{T} \tilde{v}_j|^\gamma} \\
&\rightarrow_p \infty.
\end{aligned}$$

Therefore,

$$P(j \in \mathcal{B}_T) \leq P\left(\frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{u}),j}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{u})} \hat{\mathbf{v}}) = \frac{1}{\sqrt{T}} \lambda_{(\mathbf{u})} w_{2,j}\right) \rightarrow 0.$$

The proof of the theorem is completed.

5.2 General Case

Suppose the true model is given as (1.1), where \mathbf{C} is a rank- r matrix ($r > 1$). We assume the rank of \mathbf{C} has been correctly identified, and we also assume some \sqrt{T} -consistent estimate $\tilde{\mathbf{C}}$ of \mathbf{C} , say, $\tilde{\mathbf{C}}$ is available, and $\sqrt{T} \text{vec}(\tilde{\mathbf{C}} - \mathbf{C}) \rightarrow_d N(0, \Sigma_c)$. According to the exclusive-extraction method, we estimate each unit-rank layer \mathbf{C}_k ($k = 1, \dots, r$) by minimizing the following objective function $Q(\mathbf{u}, \mathbf{v})$:

$$Q(\mathbf{u}, \mathbf{v}) = \frac{1}{2} \text{tr}\{[\mathbf{S}_k - \mathbf{u} \mathbf{v}^T \mathbf{G}][\mathbf{S}_k - \mathbf{u} \mathbf{v}^T \mathbf{G}]^T\} + \lambda_T \sum_{i=1}^m \sum_{j=1}^n w_{ij} |u_i v_j| \quad (5.3) \quad \text{comment icon}$$

where $\mathbf{S}_k = \mathbf{S} - \tilde{\mathbf{C}}_{-k} \mathbf{G}$ with $\tilde{\mathbf{C}}_{-k} = \tilde{\mathbf{C}} - \tilde{\mathbf{C}}_k$. Similarly as before, the singular value is absorbed into the singular vectors, and we assume $\mathbf{C}_k \in \Omega_1$ with $\mathbf{C}_k = \mathbf{u}^* \mathbf{v}^{*T}$ and

$u_1^* = 1$. All the other settings are the same as before.

Theorem 4. Suppose condition **C1** and **C2** are satisfied, and suppose $\frac{\lambda_T}{\sqrt{T}} \rightarrow \lambda_0 \geq 0$ as $T \rightarrow \infty$. Then there exists a local minimizer $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ of $Q_T(\mathbf{u}, \mathbf{v})$ in (5.3) such that $\|\hat{\mathbf{u}}^{(T)} - \mathbf{u}^*\| = O_p(T^{-\frac{1}{2}})$ and $\|\hat{\mathbf{v}}^{(T)} - \mathbf{v}^*\| = O_p(T^{-\frac{1}{2}})$.

Proof: We follow the proof of Theorem 1. The only difference is the expression of $\Psi_T(\mathbf{a}, \mathbf{b})$. We shall replace the term $\text{vec}(\frac{1}{\sqrt{T}} \mathbf{G} \mathbf{E}^T)$ by $\text{vec}[\frac{1}{\sqrt{T}} \mathbf{G} (\mathbf{S} - \tilde{\mathbf{C}}_{-k} \mathbf{G} - \mathbf{u}^* \mathbf{v}^{*T} \mathbf{G})^T]$. By some algebra, we have

$$\begin{aligned} & \text{vec}[\frac{1}{\sqrt{T}} \mathbf{G} (\mathbf{S} - \tilde{\mathbf{C}}_{-k} \mathbf{G} - \mathbf{u}^* \mathbf{v}^{*T} \mathbf{G})^T] \\ &= (\mathbf{I}_m \otimes \frac{\mathbf{G} \mathbf{G}^T}{T}) \sqrt{T} [\text{vec}(\mathbf{C}_{-k}) - \text{vec}(\tilde{\mathbf{C}}_{-k})] + \text{vec}(\frac{1}{\sqrt{T}} \mathbf{G} \mathbf{E}^T) \end{aligned}$$

By **C1**, **C2** and Lemma 3, the above expression is $O_p(1)$. The rest of the proof is exactly the same as the proof of Theorem 1.

Theorem 5. Suppose condition **C1-C3** are satisfied. Let $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ be the local minimizer of $Q_T(\mathbf{u}, \mathbf{v})$ in (5.3) as found in Theorem 1. Then $\sqrt{T}(\hat{\mathbf{u}}_{\mathcal{A}-1}^{(T)} - \mathbf{u}_{\mathcal{A}-1}^*)$ and $\sqrt{T}(\hat{\mathbf{v}}_{\mathcal{B}}^{(T)} - \mathbf{v}_{\mathcal{B}}^*)$ are both asymptotically normally distributed.

Proof: We follow the proof of Theorem 2. Again, the difference is the expression of $\Psi_T(\mathbf{a}, \mathbf{b})$. We shall replace the term $\text{vec}(\frac{1}{\sqrt{T}} \mathbf{G} \mathbf{E}^T)$ by

$$(\mathbf{I}_m \otimes \frac{\mathbf{G} \mathbf{G}^T}{T}) \sqrt{T} [\text{vec}(\mathbf{C}_{-k}) - \text{vec}(\tilde{\mathbf{C}}_{-k})] + \text{vec}(\frac{1}{\sqrt{T}} \mathbf{G} \mathbf{E}^T).$$

By **C1**, **C2** and Lemma 3, it can be seen that this expression is asymptotically normally distributed. We shall then modify the distribution of $\mathbf{w}_{\mathcal{AB}}$ accordingly. The rest of the proof is exactly the same as the proof of Theorem 2.

Theorem 6. Suppose condition **C1-C3** are satisfied. Let $(\hat{\mathbf{u}}^{(T)}, \hat{\mathbf{v}}^{(T)})$ be the local minimizer of $Q_T(\mathbf{u}, \mathbf{v})$ in (5.3) as found in Theorem 1. Let $\mathcal{A}_T = \{i : \hat{u}_i^{(T)} \neq 0\}$ and $\mathcal{B}_T = \{j : \hat{v}_j^{(T)} \neq 0\}$. Then $P(\mathcal{A}_T = \mathcal{A}) \rightarrow 1$ and $P(\mathcal{B}_T = \mathcal{B}) \rightarrow 1$ as $T \rightarrow \infty$.

Proof: We follow the proof of Theorem 3. In this setting, the left-hand side of (5.4)

becomes:

$$\begin{aligned}
LHS &= \frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{v}),i}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{v})} \hat{\mathbf{u}}) \\
&= \frac{1}{\sqrt{T}} [\hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \mathbf{v}^* u_i^* + \hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \left(\sum_{j \neq k}^r \mathbf{v}_j^* u_i^* - \sum_{j \neq k}^r \tilde{\mathbf{v}}_j \tilde{u}_i \right) - \hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \hat{\mathbf{v}} \hat{u}_i + \hat{\mathbf{v}}^T \mathbf{G} \mathbf{e}_{(i)}] \\
&= \hat{\mathbf{v}}^T \frac{\mathbf{G} \mathbf{G}^T}{T} \sqrt{T} \left(\sum_{j \neq k}^r \mathbf{v}_j^* u_i^* - \sum_{j \neq k}^r \tilde{\mathbf{v}}_j \tilde{u}_i \right) - \frac{\hat{\mathbf{v}}^T \mathbf{G} \mathbf{G}^T \hat{\mathbf{v}}}{T} \sqrt{T} \hat{u}_i + \frac{\hat{\mathbf{v}}^T \mathbf{G} \mathbf{e}_{(i)}}{\sqrt{T}} \\
&= O_p(1)
\end{aligned}$$

Note that this is true by the fact that $\sqrt{T} \left(\sum_{j \neq k}^r \mathbf{v}_j^* u_i^* - \sum_{j \neq k}^r \tilde{\mathbf{v}}_j \tilde{u}_i \right) = O_p(1)$ by Lemma 3. The left-hand side of 5.5 becomes:

$$\begin{aligned}
LHS &= \frac{1}{\sqrt{T}} \mathbf{X}_{(\mathbf{u}),j}^T (\mathbf{y} - \mathbf{X}_{(\mathbf{u})} \hat{\mathbf{v}}) \\
&= \frac{1}{\sqrt{T}} [\hat{\mathbf{u}}^T (\mathbf{C}_{-k}^* - \tilde{\mathbf{C}}_{-k}) \mathbf{G} \mathbf{g}_{(i)} + \hat{\mathbf{u}}^T \mathbf{u}^* \mathbf{v}^{*T} \mathbf{G} \mathbf{g}_{(i)} - \hat{\mathbf{u}}^T \hat{\mathbf{u}} \hat{\mathbf{v}}^T \mathbf{G} \mathbf{g}_{(i)} + \hat{\mathbf{u}}^T \mathbf{E} \mathbf{g}_{(i)}] \\
&= \hat{\mathbf{u}}^T \sqrt{T} (\mathbf{C}_{-k}^* - \tilde{\mathbf{C}}_{-k}) \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} \\
&\quad + \sqrt{T} (\hat{\mathbf{u}} - \mathbf{u}^*)^T \mathbf{u}^* \mathbf{v}^{*T} \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} - \hat{\mathbf{u}}^T \hat{\mathbf{u}} \sqrt{T} (\hat{\mathbf{v}}^T - \mathbf{v}^{*T}) \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} \\
&\quad - \sqrt{T} (\hat{\mathbf{u}}^T \hat{\mathbf{u}} - \mathbf{u}^{*T} \mathbf{u}^*) \mathbf{v}^{*T} \frac{\mathbf{G} \mathbf{g}_{(i)}}{T} + \frac{\hat{\mathbf{u}}^T \mathbf{E} \mathbf{g}_{(i)}}{\sqrt{T}} \\
&= O_p(1).
\end{aligned}$$

Note that this is true by the fact that $\sqrt{T} (\mathbf{C}_{-k}^* - \tilde{\mathbf{C}}_{-k}) = O_p(1)$ by Lemma 3. The rest of the proof is exactly the same as the proof of Theorem 3.

6 Discussion



- form of the penalty.
- biconvexity.
- orthogonality.

References

Akaike, H. (1974), “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, 19, 716–723.

Anderson, T. W. and Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis, 2nd Edition*, Wiley-Interscience, 2nd ed.

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,” *Proceedings of the National Academy of Sciences of the United States of America*, 98, 13790–13795.

Busygin, S., Prokopyev, O., and Pardalos, P. M. (2008), “Biclustering in data mining,” *Computers & Operations Research*, 35, 2964 – 2987, part Special Issue: Bio-inspired Methods in Combinatorial Optimization.

Chan, K.-S., Stenseth, N. C., Kittilsen, M., Gjøsæter, J., Lekve, K., and Smith, T. (2003a), “Assessing the effectiveness of releasing cod larvae for stock improvement with monitoring data,” *Ecological Applications*, 13, 3–22.

Chan, K.-S., Stenseth, N. C., Lekve, K., and Gjøsæter, J. (2003b), “Modeling Pulse Disturbance Impact On Cod Population Dynamics: The 1988 Algal Bloom of Skagerrak, Norway,” *Ecological Applications*, 73(1), 151–171.

Chen, K. and Chan, K.-S. (2010), “Testing Common Dynamics Between Exposed Fjords and Inner Fjords,” *Unpublished manuscript*.

Efron, B., Hastie, T., Johnstones, I., and Tibshirani, R. (2004), “Least Angle Regression,” *The Annals of Statistics*, 32(2), 407–499.

Friedman, J. (2007), “Pathwise Coordinate Optimization,” *The Annals of Statistics*, 1(2), 302–332.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010), “Regularization Paths for Generalized Linear Models via Coordinate Descent,” *Journal of Statistical Software*, 33, 1–22.

Gorski, J., Pfeuffer, F., and Klamroth, K. (2007), “Biconvex sets and optimization with biconvex functions: a survey and extensions,” *Mathematical Methods of Operations Research*, 66(3), 373–407.

Izenman, A. J. (1975), “Reduced-rank regression for the multivariate linear model,” *Journal of Multivariate Analysis*, 5, 248–264.

Lee, M., Shen, H., Huang, J. Z., and Marron, J. S. (2010), “Bioclustering via Sparse Singular Value Decomposition,” *Biometrics*.

Liu, Y., Hayes, D. N. N., Nobel, A., and Marron, J. S. (2008), “Statistical Significance of Clustering for High-Dimension, Low Sample Size Data,” *Journal of the American Statistical Association*, 103, 1281–1293.

Reinsel, G. C. and Velu, P. (1998), *Multivariate reduced-rank regression: theory and applications*, New York: Springer.

Schwarz, G. (1978), “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461–464.

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. 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,” *Journal of the Royal Statistical Society (Series B)*, 58, 267–288.

van der Vaart, A. W. (2000), *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*, Cambridge University Press.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009), “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis.” *Biostatistics (Oxford, England)*, 10, 515–534.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), “Dimension reduction and coefficient estimation in multivariate linear regression,” *Journal Of The Royal Statistical Society Series B*, 69, 329–346.

Zou, H. (2006), “The Adaptive Lasso and Its Oracle Properties,” *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007), “On the degree of freedom of the lasso,” *The Annals of Statistics*, 35, 2173–2192.