Joint variable and rank selection for parsimonious estimation of high dimensional matrices

Florentina Bunea*, Yiyuan She† and Marten H. Wegkamp‡

Abstract

This article is devoted to optimal dimension reduction methods for sparse, high dimensional multivariate response regression models. Both the number of responses and that of the predictors may exceed the sample size. Sometimes viewed as complementary, predictor selection and rank reduction are the most popular strategies for obtaining lower dimensional approximations of the parameter matrix in such models. We show in this article that important gains in prediction accuracy can be obtained by considering them jointly. For this, we first motivate a new class of sparse multivariate regression models, in which the coefficient matrix has low rank and zero rows or can be well approximated by such a matrix. Then, we introduce estimators that are based on penalized least squares, with novel penalties that impose simultaneous row and rank restrictions on the coefficient matrix. We prove that these estimators indeed adapt to the

*Department of Statistical Science, Cornell University
†Department of Statistics, Florida State University
‡Department of Mathematics & Department of Statistical Science, Cornell University
unknown matrix sparsity and have fast rates of convergence. We support our theoretical results with an extensive simulation study and two data analyses.

1 Introduction

The multivariate response regression model

\[ Y = XA + E \]  \hspace{1cm} (1)

has seen a revival in the last five years, with the development of computationally efficient estimation methods appropriate for high dimensional data and the emergence of accompanying finite sample theoretical analyses of the resulting estimators. The model postulates a linear relationship between \( Y \), the \( m \times n \) matrix containing measurements on \( n \) responses for \( m \) subjects and \( X \), the \( m \times p \) matrix of measurements on \( p \) predictor variables, of rank \( q \). The term \( E \) is an unobserved \( m \times n \) zero mean random noise matrix. The unknown \( p \times n \) coefficient matrix \( A \) of unknown rank \( r \) needs to be estimated. If we use (1) to model complex data sets, with a high number of responses and predictors, the number of unknowns can quickly exceed the sample size \( m \), but the situation need not be hopeless for the following reason. Let \( r(A) \) (or simply \( r \)) denote the rank of \( A \) and \( J(A) \) (or simply \( J \)) denote the index set of the non-zero rows of \( A \) and \( |J(A)| \) its cardinality. Counting the parameters in the singular value decomposition of \( A \), we observe that in fact only \( r(n + |J| - r) \) free parameters need to be estimated, and this can be substantially lower than the sample size \( m \). Furthermore, as we can always reduce \( X \) of rank \( q \) to an \( m \times q \) matrix with \( q \) independent columns in \( \mathbb{R}^m \) that span...
the same space as the columns of $X$, the corresponding coefficient matrix will have $q$ rows, so we can always assume that $|J| \leq q$. If $A$ is of full rank with no zero rows then the total number of parameters to be estimated reverts back to $nq$. If either, or both, $q$ and $n$ are large, more parsimonious models have to be proposed. Among the possible choices, two are particularly popular.

The first class is of what we will call rank sparse or rank deficient models, which postulate either that $A$ has low rank or that it can be well approximated by a low rank matrix. Interest in this class has spiked very recently, as discussed further in Section 1.1 below. Methods tailored to rank sparsity seek adaptive rank $k$ approximations of the coefficient matrix $A$. Then, one only needs to estimate $k(q + n - k)$ parameters, which can be substantially less than $nq$ for low values of $k$. This type of methods are particularly powerful if, in addition, the selected rank $k$ is close to $r$, the true rank of $A$.

The second class of models reflects the belief that $|J|$ is smaller than $q$, and we will call them row sparse models. Methods that adapt to row sparsity belong to the variable selection class, as explained in Section 1.1 below. The effective number of parameters of such models is $|J|n$. This number is smaller than the unrestricted $nq$, but may be higher than $r(|J| + n - r)$, especially if the rank of $A$ is low.

This discussion underlines the need for introducing and studying another class of models, that will embody both sparsity constraints on $A$ simultaneously. In this work we introduce row and rank sparse models, and suggest and analyze methods that combine
the strengths of the existing dimension reduction techniques. We propose a sequence
of methods that utilize penalized least squares, with new penalties that are tailored to
adaptive and optimal estimation in the row and rank sparse model (1). The rest of the
article is organized as follows.

In Section 2.1 below we introduce the product-type penalty that imposes simultane-
ously rank and row sparsity restrictions on the coefficient matrix. It generalizes the
AIC type penalties developed for variable selection in univariate response regression
models and, also, the rank penalties recently introduced by Bunea, She, and Wegkamp
(2011b) for low rank estimation in multivariate response models. The purpose of the
resulting method is two-fold. First, we prove in Theorem 1 of Section 2.1 that the
resulting estimators of $A$ adapt to both types of sparsity, row and rank, under no
conditions on the design matrix. Moreover, their rates of convergence coincide with
the existing minimax rates in the literature, cf. Koltchinskii, Lounici, and Tsybakov
(2011). Secondly, we show in Theorem 2 below that this method can also be employed
for selecting among competing estimators from a large finite list. The selected estima-
tor achieves the best bias-variance trade-off relative to the list. This is of particular
interest for selecting among estimates of different ranks and sparsity patterns, possibly
obtained via different methods. The results of Section 2.1 hold for any values of $m, n$
and $p$ and, in particular, both $n$ and $p$ can grow with $n$, but computing the estimator
analyzed in Theorem 1 requires an exhaustive search over the class of all possible mod-
els, the size of which is exponential in $p$, and this becomes computationally prohibitive
if $p > 20$. 
To address the computational issue, while still maintaining the theoretical optimality of our estimators we propose two other methods in Section 2.2. The crucial ingredient of both methods is the selection of predictors in multivariate response regression models under rank restrictions. We define and analyze this core procedure in Section 2.2, and describe a computationally efficient algorithm in Section 3.1. By combining this method with two different ways of selecting the rank adaptively we obtain two different final estimators of $A$. Both are computable in high dimensions, and both achieve the optimal rates discussed in Section 2.1, under different (mild) assumptions. We also compare the theoretical advantages of these new methods over a simple two stage procedure in which one first selects the predictors, and then reduces the rank. We illustrate the practical differences via a simulation study in Section 3.2. We then use our methods for the analysis, presented in Section 4, of two data sets arising in machine learning and cognitive neuroscience, respectively. The proofs of our results are collected in the Appendix.

1.1 Background

Before we introduce and discuss our methods, we give an overview of existing procedures of adaptive estimation in (1), that adapt to either rank or row sparsity, but not both. We also present a comparison of target rates under various sparsity assumptions on the coefficient matrix $A$ in model (1).
Reduced rank estimation of $A$ in (1) and the immediate extensions to canonical correlation analysis (CCA) and principal components analysis (PCA) are perhaps the most popular ways of achieving dimension reduction of multivariate data, and they have become a standard tool in time series (Brillinger, 1981), econometrics (Reinsel and Velu, 1998) and machine learning (Izenman, 2008), to name just a few areas. The literature on low rank regression estimation of $A$ dates back to Anderson (1951). The model is perhaps best known as \textit{reduced-rank regression} RRR (Izenman, 2008) and, until recently, it had only been studied theoretically from an asymptotic perspective, in a large sample size regime. We refer to Reinsel and Velu (1998) for a historical development and references, and to Izenman (2008) for a large number of applications and extensions. Very recently, a number of works proposed and analyzed adaptive methods that yield rate optimal low rank approximations of $A$ in high dimensions, with estimators computable even if $n$ and/or $p$ exceed the sample size $m$. These estimators minimize over all matrices $B$ the penalized square Frobenius norm $\|Y - XB\|_F^2$, where $\|M\|_F^2$ denotes the sum of the squared entries of a generic matrix $M$. For penalties proportional to the nuclear norm of $B$, which is the sum of the singular values of $B$, we refer to Yuan, Ekici, Lu, and Monteiro (2007), Candès and Plan (2010), Negahban and Wainwright (2011), Rohde and Tsybakov (2011). For penalties proportional to the rank of $B$ we refer to Bunea, She, and Wegkamp (2011b) and Giraud (2011). Both types of estimators are computationally efficient. They achieve, adaptively, the rate of convergence $(q + n)r$ which, under suitable regularity conditions is the optimal minimax rate in (1) under rank sparsity, see, e.g. Rohde and Tsybakov (2011) for lower bound calculations.
To explain the other notion of sparsity, note first that postulating that predictor $X_j$ is not in model (1) is equivalent with assuming that the $j$-th row in $A$ is identically zero. Notice further that after applying the vectorization operator to both sides of model (1) we obtain a univariate response regression model. Then, the rows of $A$ can be regarded as groups of coefficients in the transformed model, and therefore can be set to zero by any group selection method developed for univariate response regression models. Perhaps one of the most popular and best understood group selection methods in high dimensions is the Group Lasso (Yuan and Lin, 2006), GLASSO for later reference, and we will also use it for comparison in this work. The optimal minimax rate in (1) under row sparsity is proportional to $|J|n + |J| \log(p/|J|)$, again under suitable regularity conditions, see Lounici, Pontil, Tsybakov, and van de Geer (2010) and Wei and Huang (2010).

Despite these very recent advances, adaptive low rank estimation in (1), based on a reduced set of predictors, has not been investigated either theoretically or practically. Previous attempts to achieve a similar goal belong mainly to the class of sparse PCA algorithms, see, e.g., SCotLass (Jolliffe, Trendafilov, and Uddin, 2003), sparse PCA (Zou, Hastie, and Tibshirani, 2006; Shen and Huang, 2008), and the penalized matrix decomposition (PMD) (Witten, Tibshirani, and Hastie, 2009), among others. These algorithms are tailored to the PCA model, and are not easily extendable to the general model (1). Moreover, although each of the resulting principal components generated by these procedures are sparse, they are linear combinations of possibly different subsets of the original $p$ variables. It is therefore possible that all $p$ variables
or a large subset of them will still be needed for the final analysis. The methods we propose in this work address these potential drawbacks, are directly developed for the general model (1) and come with theoretical guarantees. For ease of reference, Table 1 below contains a rate comparison between optimal prediction error rates achievable by variable selection (GLASSO), low rank estimation (RSC and NNP) and our new joint rank and row selection (JRRS) methods, respectively.

Table 1: Oracle rate comparison between JRRS, GLASSO, and RSC. The sample size and dimension parameters \(m, p, n, q, r, |J|\) satisfy \(q \leq m \land p, r \leq n \land |J|, |J| \leq q\).

<table>
<thead>
<tr>
<th>Method</th>
<th>Rate Comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLASSO</td>
<td>(</td>
</tr>
<tr>
<td>RSC or NNP</td>
<td>((q + n)r)</td>
</tr>
<tr>
<td>JRRS</td>
<td>((</td>
</tr>
</tbody>
</table>

Since we always have \((n + |J|)r \leq 2(n \lor |J|)(n \land |J|) \leq 2|J|n\), irrespective of the values of the parameters involved, it is clear that further imposing rank constraints in the estimation procedure always improves the estimation error rate, over a pure variable selection type estimation. Therefore the proposed JRRS-type estimators always have better rates than the GLASSO, sometimes much better, as the inequality above is blunt. Clearly, adding the row sparsity to the rank sparsity constraints always improves the rate over low rank estimation and, again, JRRS-type estimation improves over either RSC or NNP. The table also reveals an important fact: if \(n > q\), the rates of the RSC, NNP and JRRS are dominated by \(nr\), regardless of \(J\). If the number of responses \(n\) is smaller than \(q\), then the new class of methods can provide substantial rate improvements over the existing methods.
2 Adaptation to row and rank sparsity: estimation procedures and oracle inequalities

2.1 The single-stage joint rank and row selection estimator

In this section we modify the rank selection criterion (RSC) introduced in Bunea, She, and Wegkamp (2011b) to accommodate variable selection. Our new criterion is the penalized least squares criterion \( \|Y - XB\|^2_F + \text{pen}(B) \) with penalty term

\[
\text{pen}(B) = c\sigma^2(2n + |J(B)|)r(B)
\]

that is essentially proportional to the number of parameters in a model with fewer predictors \( J(B) \) and of reduced rank \( r(B) \). Here \( c > 1 \) is a numerical constant, \( J(B) \) is the set of indices of non-zero rows, \( r(B) \) is rank of a generic \( p \times n \) matrix \( B \) and the Frobenius norm of a generic matrix \( M \) is denoted by \( \|M\|^2_F \) and is equal to the sum of the squared entries of \( M \). We define

\[
\hat{B} = \arg \min_B \{\|Y - XB\|^2_F + \text{pen}(B)\}
\]

as our single-stage Joint Rank and Row Selection (JRRS) estimator, denoted by \( \text{JRSSL} \). If \( \hat{B} \) is computed by minimizing over all \( p \times n \) matrices \( B \), then Theorem 1 stated below shows that it adapts optimally to the unknown row and rank sparsity of \( A \): the mean squared error of \( \hat{B} \) coincides with that of optimal estimators of rank \( r \) and with \( |J| \) non-zero rows, had these values been known prior to estimation. However, the construction
of $\hat{B}$ does not utilize knowledge of either $r$ or $J$, hence the term adaptive. The minimax lower bounds for this model can be obtained by an immediate modification of Theorem 5 in Koltchinskii, Lounici, and Tsybakov (2011). Our single-stage JRRS estimator $\hat{B}$ given in (3) above achieves the lower bound, under no restrictions on the design $X$, rank $r$, or dimensions $m, n, p$.

**Theorem 1.** For any $c > 1$, the single-stage JRRS estimator $\hat{B}$ in (3) satisfies

$$\mathbb{E} \left[ \| XA - X\hat{B} \|_F^2 \right] \lesssim \inf_{\hat{B}} \left[ \| XA - XB \|_F^2 + \sigma^2 (n + |J(B)|) r(B) \right]$$

$$\lesssim \sigma^2 \{ n + |J(A)| \} r(A).$$

*Here and elsewhere $\lesssim$ means that the inequality holds up to multiplicative numerical constants.*

The first inequality shows that our estimator achieves the optimal “bias-variance” trade-off over all matrices $B$. The first term of this inequality quantifies the (squared) bias and the second term, proportional to the number of parameters in a model with $J(B)$ predictors and of rank $r(B)$, plays the role of the variance component. Then, in particular, if $B = A$, we obtain the second inequality, which shows that our estimator is minimax rate optimal.

As we discussed in the introduction, although the procedure leading to $\hat{B}$ can be used for selecting the best among all $p \times n$ matrices $B$, it may become computationally intractable when $p > 20$, as we need to consider all $2^p$ subsets of predictors in our search. However, and very importantly, the proof of Theorem 1 is valid even if the
matrices we select from are random and/or depend on the data, which means that our procedure can be used for selecting from any countable list of random matrices of different ranks and with different sparsity patterns. To the best of our knowledge, this is the first non-asymptotic procedure with this feature, and we state it separately below to re-emphasize the usage of the procedure and the fact that the selected estimator achieves the best bias-variance trade-off relative to all the matrices on the list. We will make essential use of this result in the next section.

**Theorem 2.** For any collection of (random) matrices \( B_1, B_2, \ldots \), the single-stage JRRS estimator

\[
\tilde{B} = \arg \min_{B_j} \left\{ \| Y - X B_j \|_F^2 + c \sigma^2 (2n + |J(B_j)|) r(B_j) \right\}
\]

for any \( c > 1 \), satisfies

\[
\mathbb{E} \left[ \| XA - XB \|_F^2 \right] \leq \inf_j \mathbb{E} \left[ \| XA - XB_j \|_F^2 + \sigma^2 (n + |J(B_j)|) r(B_j) \right].
\]

### 2.2 Two-step joint rank and row selection estimators

The computational complexity of the single-stage JRRS estimator (3) is owed to the component of the penalty term proportional to \( J(B) \), which is responsible for row selection. The existence of this term in (2) forces complete enumeration of the model space. We address this problem by proposing a convex relaxation \( \| B \|_{2,1} \) of this component. Here \( \| B \|_{2,1} = \sum_{j=1}^p \| b_j \|_2 \) is the sum of the Euclidean norms \( \| b_j \|_2 \) of the rows \( b_j \) of \( B \). In this section, we propose two alternative, two-step JRRS procedures, each building
on the following core estimator.

2.2.1 Rank-constrained predictor selection

We define our rank-constrained row-sparse estimators $\hat{B}_k$ of $A$ as

$$\hat{B}_k = \arg \min_{\text{rank}(B) \leq k} \left\{ \frac{1}{2} \|Y - XB\|_F^2 + \lambda \|B\|_{2,1} \right\}. \quad (5)$$

Here $\lambda$ is a tuning parameter that we discuss below and the minimization is over all $p \times n$ matrices $B$ of rank less than or equal to (a fixed) $k$. We give a computationally efficient numerical algorithm for solving this minimization problem in Section 3.1 below. Clearly, for $k = q$, there is no rank restriction in (5) and the resulting estimator is the GLASSO estimator; for $\lambda = 0$, we obtain the reduced-rank regression estimator. Thus, the procedure yielding the estimators $\hat{B}_k$ of rank $k$ acts as a synthesis of the two dimension reduction strategies, having each of them as limiting points. We will refer to the estimators $\hat{B}_k$ as the Rank Constrained Group Lasso (RCGL) estimators.

Since this estimator is central to our procedures developed in Section 2.2.2 below, we begin by analyzing its performance. In Theorem 3, we prove that $\hat{B}_k$ adapts to row sparsity, for each $k$, under the following additional mild assumption on the design matrix $X$. Let $\Sigma = X'X/m$ be the standardized Gram matrix.

Assumption 2. We say $\Sigma \in \mathbb{R}^{p \times p}$ satisfies condition $\mathcal{A}(I, \delta_I)$ for an index set $I \subseteq \mathbb{R}$.
\{1, \ldots, p\} and a positive number \( \delta_I \), iff

\[ a'^* \Sigma a \geq \delta_I \sum_{j \in I} a_j^2 \quad \text{for any } a \in \mathbb{R}^p. \] (6)

Assumption \( \mathcal{A} \) on \( \Sigma \) essentially requires that only a sub-matrix of \( \Sigma \) has a non-zero smallest eigenvalue, allowing therefore designs with \( p > m \). Since \( \Sigma \) is always positive semidefinite, \( \mathcal{A} \) only postulates that it remains that way when we subtract very small quantities from few of its diagonal elements and leave the rest unchanged. This can be regarded as a stability requirement on \( \Sigma \) and is very mild indeed.

In what follows we establish the theoretical properties of \( \hat{B}_k \). Let \( \rho = \lambda_1(\Sigma) \) denote the largest eigen-value of the standardized Gram matrix \( \Sigma \) and set our tuning parameter \( \lambda \) as

\[ \lambda = C \sigma \sqrt{\rho km}, \] (7)

for some numerical constant \( C > 0 \) large enough. Notice that \( \lambda \) depends on \( k \), but we suppress this dependence in our notation.

**Theorem 3.** Let \( \hat{B}_k \) be a global minimizer of (5) corresponding to \( \lambda \) in (7). Then, we have

\[ \mathbb{E} \left[ \|X\hat{B}_k - XA\|_F^2 \right] \leq \|XB - XA\|_F^2 + \left\{ n + |J(B)| \left( 1 + \frac{\rho}{\delta_{J(B)}} \right) \right\} k\sigma^2 \] (8)

for any \( p \times n \) matrix \( B \) with \( r(B) \leq k \), \( |J(B)| \) non-zero rows, provided \( \Sigma \) satisfies
Assumption $\mathfrak{A}(J(B), \delta_{J(B)})$.

Notice that the term $\{n + |J(B)|\}k\sigma^2$ in (8) is multiplied by a factor $1 + \{\rho/\delta_{J(B)}\}$. If this factor stays bounded, then Theorem 3 shows that, within the class of row sparse matrices of fixed rank $k$, the RCGL estimator is row-sparsity adaptive, in that the best number of predictors does not have to be specified prior to estimation.

Moreover, if one knew the rank $r$ of $A$ in advance, then the corollary below shows that RCGL achieves the desired rate of convergence in row and rank sparse models, adaptively over rows (and non-adaptively over ranks).

The factor $\kappa = \rho/\delta_{J(A)}$ can be viewed as a generalized condition number of the (standardized) Gram matrix $\Sigma = X'X/m$. For complicated, highly correlated designs, the statistical problem is more challenging than for near orthogonal designs, and this is reflected by the increase of the rate by the factor $\kappa$.

It is interesting to contrast our estimator with the regular GLASSO estimator $\hat{A}$ that minimizes $\|Y - XB\|_F^2 + 2\lambda\|B\|_{2,1}$ over all $p \times n$ matrices $B$. Our choice (7) of the tuning parameter $\lambda$ markedly differs from the choice

$$\lambda' = 2\sqrt{2}\sigma\sqrt{mn}\left(1 + \frac{K\log p}{n}\right)^{1/2}$$

with $K > 5/2$, proposed by Lounici, Pontil, Tsybakov, and van de Geer (2010) for the GLASSO estimator $\hat{A}$. These authors obtain the rate $\{J\}n\sigma^2\delta^{-1} \{1 + (K \log p)/n\}$ for the squared Frobenius norm $\|XA - X\hat{A}\|_F^2$ of the resulting GLASSO estimator $\hat{A}$ and show that this rate is close to minimax optimal. This rate, however, is too large for
our needs. Our choice \( \lambda = C\sigma \sqrt{mk\sqrt{\rho}} \) replaces \( n \) by \( k \) (and we recall that \( k \leq n \)), reflecting the fact that we minimize in (5) over all rank-\( k \) matrices, not all \( p \times n \) matrices. We need a more refined analysis than the one for the GLASSO to derive the desired bound \( (n + |J|)k\sigma^2 \). We impose condition \( \mathfrak{A} \) that is slightly stronger than condition 3.1 in Lounici, Pontil, Tsybakov, and van de Geer (2010), who require (6) to hold over a subset of all \( p \times n \) matrices \( M \) only. Finally, we do not have the superfluous log term appearing in the penalty and rate - instead we have the term \( \rho = \lambda_1(\Sigma) \).

**Corollary 4.** Under the assumptions of Theorem 3, for \( k = r = r(A) \),

\[
\mathbb{E} \left[ \|X\widehat{B}_r - X\Lambda\|_F^2 \right] \lesssim (n + |J|) r\sigma^2
\]

(9)

provided \( \Sigma \) satisfies Assumption \( \mathfrak{A}(J,\delta_J) \), and \( \rho/\delta_J \) is of order \( O(1) \).

### 2.2.2 Adaptive rank-constrained predictor selection

In the following, we will investigate theoretical properties of three methods, **Method 1** (RSC→RCGL), **Method 2** (RCGL→JRRS1), and **Method 3** (GLASSO→RSC), where JRRS1 denotes the single-stage JRRS estimator of Section 2.1.

First, Corollary 4 suggests that complementing RCGL by a method that estimates the rank consistently we could obtain row and rank optimal adaptive estimator. This is indeed true, and motivates the procedure described below for obtaining our final estimator.

**Method 1 (RSC→RCGL)**
• **Step 1.** Use the Rank Selection Criterion (RSC) proposed and analyzed in Bunea, She, and Wegkamp (2011b) to select (consistently) \( k = \hat{r} \).

• **Step 2.** Compute the rank constrained GLASSO estimator \( \hat{B}_k \) in (5) above with \( k = \hat{r} \) to obtain the final estimator \( \hat{B}^{(1)} = \hat{B}_\hat{r} \).

For step 1, we use Proposition 1 in Bunea, She, and Wegkamp (2011b) to select the rank \( \hat{r} \) as the number of singular values \( d_k(PY) \) of \( PY \) that exceed \( \sqrt{\mu} \equiv \sigma(\sqrt{2n} + \sqrt{2q}) \) and then appeal to Theorem 2 in the same work to guarantee the consistency of the procedure, under minimal conditions. Here \( P = X(X'X)^{-1}X' \) is the projection matrix onto the space spanned by \( X \). We show below that this two-step estimator adapts to both rank and row sparsity, under two additional, mild restrictions.

**Assumption C1.** \( d_r(XA) > 2\sigma(\sqrt{n} + \sqrt{q}) \)

**Assumption C2.** \( \log(\|XA\|_F) \leq (\sqrt{2} - 1)^2(n + q)/4 \).

Assumption **C1** only requires that the signal strength, measured by \( d_r(XA) \), be larger than the “noise level” \( 2\sigma\sqrt{2n} + 2q \), otherwise its detection would become problematic. The tightness of **C1** is discussed in detail in Bunea, She, and Wegkamp (2011b). Theorem 2 of that work proves that the correct rank will be selected with overwhelming probability \( 1 - \exp\{-c_0(n + q)\} \) with \( c_0 = (\sqrt{2} - 1)^2/2 \).

The second condition, **Assumption C2**, is satisfied as soon as \( \log(mn\|XA\|_\infty) < n + q \), denoting by \( \|XA\|_\infty \) the largest entry in \( XA \), in absolute value. If the entries of \( XA \) are uniformly bounded, then **C2** essentially postulates that the sample size \( m \) be smaller
than \( \exp(n + q) \). In fact, we may weaken \( \mathfrak{C}_2 \) a bit by requiring the bound

\[
\log(\|XA\|_F) \leq c_0(n + q)/2 + C\{\log(n + |J|) + \log(r)\}
\]

for some numerical constant \( C \). The proof of Theorem 5 invokes this restriction to guarantee that the error due to selecting the rank is negligible compared to the rate \( (n + |J|)r \). Our proof also shows that this (technical) condition can be dispensed with if we want to show that the inequality \( \|X\hat{B}^{(1)} - XA\|_F^2 \lesssim (n + |J|)r\sigma^2 \) holds with high probability, not in expectation.

**Theorem 5.** Let \( \Sigma \) satisfy \( \mathfrak{A}(J, \delta_J) \) with \( J = J(A) \), let \( \rho/\delta_J \) be bounded, and let \( \mathfrak{C}_1 \) and \( \mathfrak{C}_2 \) hold. \( \lambda \) is set according to (7) with \( C \) large enough. Then the two-step JRRS estimator \( \hat{B}^{(1)} \) satisfies

\[
\mathbb{E} \left[ \|X\hat{B}^{(1)} - XA\|_F^2 \right] \lesssim (n + |J|)r\sigma^2.
\]

(10)

Theorem 5 shows that \( \hat{B}^{(1)} \) is *row and rank* adaptive, and achieves the same optimal rate as the *row and rank* adaptive \( \hat{B} \) studied in Theorem 1 above, proved under no restrictions on the design. We view the minimal conditions of Theorem 5 as a small price to pay for the computational efficiency of \( \hat{B}^{(1)} \) relative to that of \( \hat{B} \) in (3) and that \( \hat{B}^{(2)} \) which will be defined in (11) below. Moreover, and very importantly, the practical choice of the threshold \( 2\sigma\sqrt{2n + 2q} \) in the initial step of our procedure can be done either by replacing \( \sigma^2 \) by an estimator, as suggested and analyzed theoretically in Section 2.4 of Bunea, She, and Wegkamp (2011b), or by cross-validation. The latter
is valid in this context for consistent rank selection, as the minimum squared error of rank restricted estimators in (1) is achieved for the true rank, as discussed in detail in Bunea, She, and Wegkamp (2011b).

In general, for adaptive row and rank adaptive estimation in model (1) there will always be a trade-off between the computational complexity of a method and the conditions that guarantee theoretical performance bounds. Indeed, we now present an alternative adaptive method that is more computationally involved than method 1, as it involves a search over a two-dimensional grid, but its analysis does not require $C_1$ and $C_2$.

**Method 2 (RCGL→JRRS)**

- **Step 1.** Pre-specify a grid $\Lambda$ of values for $\lambda$ and use (5) to construct the class $\mathcal{B} = \{\hat{B}_{k,\lambda} : 1 \leq k \leq q, \lambda \in \Lambda\}$.

- **Step 2.** Compute

$$\hat{B}^{(2)} = \arg\min_{B \in \mathcal{B}} \{\|Y - XB\|_F^2 + \text{pen}(B)\}, \quad (11)$$

with $\text{pen}(B)$ defined in (2) above.

**Theorem 6.** We have

$$\mathbb{E}[\|X\hat{B}^{(2)} - XA\|_F^2] \leq \min_{1 \leq k \leq q, \lambda \in \Lambda} \mathbb{E}\left[\|X\hat{B}_{\lambda,k} - XA\|_F^2 + (n + |J(\hat{B}_{\lambda,k})|)r(\hat{B}_{\lambda,k})\right].$$

Provided $\Sigma$ satisfies condition $\mathcal{A}(J, \delta_J)$ with $J = J(A)$, and $\rho/\delta_J = O(1)$, and $\Lambda$ con-
tains \( \lambda \) in (7) for some \( C \) large enough, we have

\[
\mathbb{E}[\|X\widehat{B}^{(2)} - XA\|_F^2] \lesssim (|J| + n)r\sigma^2.
\]

We see that \( \widehat{B}^{(2)} \) is row and rank adaptive and has the same rate as the single-stage JRRS estimator \( \widehat{B} \) under the minimal condition \( \mathfrak{A} \) on the design. Our proof, presented in the Appendix, makes use of two crucial facts: (1) \( \widehat{B}_{k,\lambda} \) adapts to row sparsity, for each \( k \), as shown in Theorem 3; and (2) \( \text{pen}(B) \) was designed for selecting estimators that achieve the best bias-variance trade-off relative to a list of possible candidates, as we showed in Theorem 2 above. Our simulation studies in Section 4 indicate that \textbf{method 2} yields numerical results that are comparable to those obtained via \textbf{method 1}.

**Remark.** Theorem 5 shows that optimal rank and row adaptivity can be achieved by following the two steps of \textbf{method 1}, under the mild assumption \( \mathfrak{A} \). Interestingly, more restrictive assumptions on the design may be needed if one used instead another, and perhaps more canonical, two-stage procedure, to which we will refer in the sequel as \textbf{Method 3 (GLASSO→RSC)}: (1) First select the predictors via the GLASSO (2) Second, use the Rank Selection Criterion (RSC) of Bunea, She, and Wegkamp (2011b) to construct an adaptive estimator, of reduced rank, based only on the selected predictors. It is clear that as soon as we select the predictors consistently in the first step, selecting consistently the rank in the second step and then proving row and rank sparsity of the resulting estimator will follow straightforward from existing
results, for instance Theorem 7 in Bunea, She, and Wegkamp (2011b). Although this is a natural path to follow, there is an important caveat to consider: the sufficient conditions under which this two-step process yields adaptive (to row and rank sparsity) estimators include the conditions under which the GLASSO yields consistent group selection. These conditions are in the spirit of those given in Bunea (2008), for the Lasso, and involve the so called mutual coherence condition on \( \Sigma = X'X/m =: (\Sigma_{ij})_{1 \leq i,j \leq p} \), which postulates that the off-diagonal elements of \( \Sigma \) be small. Specifically, for the GLASSO, the restriction becomes either \(|\Sigma_{ij}| \leq 1/(7\alpha |J|)\), for some \( \alpha > 1 \), cf. Lounici, Pontil, Tsybakov, and van de Geer (2009), if it is coupled with the condition that \( n^{-1/2}||a_j|| \geq C m^{-1/2} [1 + An^{-1/2} \log p]^{1/2} \). Here \( ||a_j|| \) is the Euclidean norm of the \( j \)th row vector of \( A \). For designs for which \( \Sigma \) is even closer to the identity matrix, in that \( |\Sigma_{ij}| \leq 1/(14\alpha |J|n) \), the condition on the minimum size of detectable coefficients can be relaxed to \( n^{-1/2}||a_j|| \geq C m^{-1/2} [1 + An^{-1} \log p]^{1/2} \), see Corollary 5.2 in Lounici, Pontil, Tsybakov, and van de Geer (2010). Our Theorems 5 and 6 require substantially weaker assumptions on the design, in particular \( \Sigma \) is only required to contain a submatrix with positive smallest eigenvalue. Moreover, accurately tuning the regularity parameter of the GLASSO algorithm for the purpose of consistent selection is often difficult in practice.
3 Computational issues and numerical performance comparison

3.1 A computational algorithm for the RCGL-estimator

In this section, we design an algorithm for minimizing

$$F(B; \lambda) := \frac{1}{2} \|Y - XB\|_F^2 + \lambda \|B\|_{2,1}$$

(12)

over all $p \times n$ matrices $B$ of rank less than or equal to $k$. We recall that by solving this problem we provide a way of performing rank-constrained variable selection in model (1). Solving directly the nonconvex constrained minimization problem for $B$ in (12) may be difficult. One way of surmounting this difficulty is to write $B = SV'$, with $V$ being orthogonal. Then the rank constrained group lasso (RCGL) optimization problem is equivalent to finding

$$(\hat{S}, \hat{V}) = \arg \min_{S \in \mathbb{R}^{p \times k}, V \in \mathcal{O}^{n \times k}} \frac{1}{2} \|Y - XSV'\|_F^2 + \lambda \|S\|_{2,1},$$

(13)

where the minimum is taken over all orthogonal $n \times k$ matrices $V$ and all $p \times k$ matrices $S$. With a slight abuse of notation, we still denote the objective function in (13) by $F(S, V; \lambda)$. We propose an iterative optimization procedure as follows.
Algorithm $\mathcal{A}$  

Rank Constrained Group Lasso (RCGL) Computation

given $1 \leq k \leq m \land p \land n$, $\lambda \geq 0$, $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n \times k}$ (say the first $k$ columns of $I_{n \times n}$ or the first $k$ right singular vectors of $Y$)

$j \leftarrow 0$, converged $\leftarrow$ FALSE

while not converged do

(a) $S_{k,\lambda}^{(j+1)} \leftarrow \arg \min_{S \in \mathbb{R}^{p \times k}} \frac{1}{2} \|YV_{k,\lambda}^{(j)} - XS\|_F^2 + \lambda \|S\|_{2,1}$.

(b) Let $W \leftarrow Y'XS_{k,\lambda}^{(j+1)} \in \mathbb{R}^{n \times k}$ and perform SVD: $W = U_wD_wV_w'$.

(c) $V_{k,\lambda}^{(j+1)} \leftarrow U_wV_w'$

(d) $B_{k,\lambda}^{(j+1)} \leftarrow S_{k,\lambda}^{(j+1)}(V_{k,\lambda}^{(j+1)})'$

(e) converged $\leftarrow |F(B_{k,\lambda}^{(j+1)}; \lambda) - F(B_{k,\lambda}^{(j)}; \lambda)| < \varepsilon$

(f) $j \leftarrow j + 1$

end while

deliver $\hat{B}_{k,\lambda} = B_{k,\lambda}^{(j+1)}$, $\hat{S}_{k,\lambda} = S_{k,\lambda}^{(j+1)}$, $\hat{V}_{k,\lambda} = V_{k,\lambda}^{(j+1)}$.

The following theorem presents a global convergence analysis for Algorithm $\mathcal{A}$, where global in this context refers to the fact that the algorithm converges for any initial point. The proof is given in the Appendix.

**Theorem 7.** Given $\lambda > 0$ and an arbitrary starting point $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n \times k}$, let $(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ $(j = 1, 2, \cdots)$ be the sequence of iterates generated by Algorithm $\mathcal{A}$. The following two statements hold:

(i) Any accumulation point of $(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ is a stationary point of $F$ and $F(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ converges monotonically to $F(S_{k,\lambda}^{*}, V_{k,\lambda}^{*})$ for some stationary point $(S_{k,\lambda}^{*}, V_{k,\lambda}^{*})$. 

22
(ii) Suppose for any \((S, V)\) outside the local minimum set of \(F\), \(\min_{\tilde{S} \in \mathbb{R}^{p \times k}} F(\tilde{S}, V) < F(S, V)\). Then, any accumulation point of \((S_{(j)}^{(k)}, V_{(j)})\) is a local minimum of \(F\) and \(F(S_{(j)}^{(k)}, V_{(j)})\) converges monotonically to \(F(S_{(k)}^{*}, V_{(k)}^{*})\) for some local minimizer \((S_{(k)}^{*}, V_{(k)}^{*})\).

Remarks.

1. To obtain the final estimate, we run the algorithm to obtain a solution path, for each \((k, \lambda)\) in a two-dimensional grid, or for a grid of \(\lambda\) with \(k\) determined by RSC. From the solution path, we get a series of candidate estimators. Then one may use the one step JRRS estimator (4) or other tuning criteria to select the optimal one.

2. Our results are of the same type as the convergence analysis for the EM algorithm (Wu, 1983). \(\mathcal{A}\) can be viewed as a block coordinate descent method, but the conclusion in Theorem 7 is stronger in some sense: the guaranteed convergence to a stationary point, which will be defined in Appendix 4.5, does not require the uniqueness of \(S_{(k+1)}^{(j)}\) in Step (a), which is a crucial assumption in the literature, see, for instance, Bertsekas (1999) and Tseng (2001). See Appendix 4.5 for the details of the proof.

3. Step (a) solves a GLASSO optimization problem. To see this, observe that

\[
\|YV_k - XS_k\|_F^2/2 + \lambda\|S_k\|_{2,1} = \|\text{vec}((YV_k)' - (X \otimes I_k)\text{vec}(S_{k}^{'})\|_F^2/2 + \lambda\|S_k\|_{2,1},
\]

where \(\text{vec}\) is the standard vectorization operator, and \(\otimes\) is the Kronecker product. Although this subproblem is convex, finding a global minimum point can still be computationally expensive for large data. One can carry out some low-cost mul-
tiplication and thresholding operations instead. Let $K$ be a constant satisfying $K > \frac{\|X\|^2}{2}$. Given any $V \in \mathbb{O}^{p \times k}$, define $T_V : \mathbb{R}^{p \times k} \rightarrow \mathbb{R}^{p \times k}$ as

$$T_V \circ S = \tilde{\Theta} \left( \frac{1}{K} X' Y V + (I - \frac{1}{K} X' X) S; \frac{\lambda}{K} \right), \forall S \in \mathbb{R}^{p \times k},$$  \hspace{1cm} (14)

where $\tilde{\Theta}$ is a multivariate version of the soft-thresholding operator $\Theta$: for any vector $a \in \mathbb{R}^k$, $\tilde{\Theta}(a; \lambda) = a \Theta(\|a\|_2; \lambda)/\|a\|_2$ for $a \neq 0$ and 0 otherwise; for any matrix $A \in \mathbb{R}^{p \times k}$ with $A = [a_1 \cdots a_p]'$, $\tilde{\Theta}(A; \lambda) = [\tilde{\Theta}(a_1; \lambda) \cdots \tilde{\Theta}(a_p; \lambda)]'$.

We now replace Step (a) in $\mathcal{A}$ by $S_{k,\lambda}^{(j+1)} \leftarrow T_{V_{k,\lambda}^{(j)}} \circ \cdots \circ T_{V_{k,\lambda}^{(j)}} \circ S_{k,\lambda}^{(j)}$, where the number of $T_{V_{k,\lambda}^{(j)}}$, denoted by $\alpha_j$, satisfies $1 \leq \alpha_j \leq M_{\text{iter}}$ for some $M_{\text{iter}} < \infty$ specified based on available computational resources. $\alpha_j$ need not be equal. This algorithm, denoted by $\mathcal{A}'$, offers more flexibility and is more convenient than $\mathcal{A}$ in implementation. Although at each iteration $S_{k,\lambda}^{(j+1)}$ is not uniquely determined, a stronger global convergence result holds for $\mathcal{A}'$. We state it below and prove it in the appendix.

**Theorem 8.** Given $\lambda > 0$ and an arbitrary starting point $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n \times k}$, let $(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)}) (j = 1, 2, \cdots)$ be the sequence of iterates generated by $\mathcal{A}'$. Then, any accumulation point of $(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ is a coordinatewise minimum point (and a stationary point) of $F$ and $F(S_{k,\lambda}^{*}, V_{k,\lambda}^{*})$ converges monotonically to $F(S_{k,\lambda}^{*}, V_{k,\lambda}^{*})$ for some coordinatewise minimum point $(S_{k,\lambda}^{*}, V_{k,\lambda}^{*})$.

### 3.2 Simulation Studies

The setup of our simulations is as follows.
• The design matrix $X$ has i.i.d. rows $X_i$ from a multivariate normal distribution $\text{MVN}(0, \Sigma)$, with \( \Sigma_{jk} = \rho^{|j-k|}, \rho > 0, 1 \leq j, k \leq p \).

• The coefficient matrix $A$ has the form

$$A = \begin{bmatrix}
A_1 \\
O
\end{bmatrix} = \begin{bmatrix}
bB_0B_1 \\
O
\end{bmatrix},$$

with $b > 0$, $B_0$ a $J \times r$ matrix and $B_1$ a $r \times n$ matrix. All entries in $B_0$ and $B_1$ are i.i.d. $N(0, 1)$.

• The noise matrix $E$ has independent $N(0, 1)$ entries. Let $E_i$ denote its $i$-th row.

• Each row $Y_i$ in $Y$ is then generated as $Y_i = X_i'A + E_i, 1 \leq i \leq m$.

This setup contains many noisy features, but the relevant features lie in a low-dimensional subspace. This structure resembles many real world datasets; see our examples in Section 4, where the low rank structure is inherent and thus rank-constrained variable selection is desired.

We report two settings:

$p > m$ setup: $m = 30$, $|J| = 15$, $p = 100$, $n = 10$, $r = 2$, $\rho = 0.1$, $\sigma^2 = 1$, $b = 0.5, 1$.

$m > p$ setup: $m = 100$, $|J| = 15$, $p = 25$, $n = 25$, $r = 5$, $\rho = 0.1$, $\sigma^2 = 1$, $b = 0.2, 0.4$.

Although we performed experiments in many other settings with $\rho = 0.5$, we do not report all results, as the conclusions are similar. The current setups show that variable
selection, without taking the rank information into consideration, may be sub-optimal even if the correlations between predictors are low.

We tested five methods: RSC, GLASSO, Method 1 (RSC→RCGL), Method 2 (RCGL→JRRS1), and Method 3 (GLASSO→RSC) as described in Section 2.2. To minimize the influence of various parameter tuning strategies on our performance comparison, we generated a large validation dataset (10,000 observations) to tune the parameter of each algorithm (with the exception of method 2) and another independent dataset of the same size as test data to evaluate the test error. Similar to the LARS-OLS hybrid (Efron, Hastie, Johnstone, and Tibshirani, 2004), for each GLASSO and RCGL estimate, we computed the least squares estimate restricted to the selected dimensions. We found that the resulting (bias corrected) solution paths are more suitable for parameter tuning. For method 2, after getting the (bias corrected) solution path, we set $c = 2$ and $\sigma^2 = 1$ in (2) to select the optimal $\hat{B}^{(2)}$; in contrast to the other two methods, no validation data is used for tuning.

Each model was simulated 50 times, and Tables 2 and 3 summarize our findings. We evaluated the prediction accuracy of each estimator $\hat{A}$ by the mean squared error (MSE) $\|XA - X\hat{A}\|^2_F/(mn)$ using the test data at each run. Since the MSE histograms turned out to be highly asymmetric, we computed the 40% trimmed-mean of MSEs as the goodness of fit of the obtained model. This trimmed mean is more robust than the mean and more stable than the median (Huber, 1981), and it therefore allows for a more fair comparison between methods.
We also report the median number of predictors (denoted by $|\widehat{J}|$) and median rank estimate (denoted by $\widehat{R}$) over all runs. Estimators with small MSE and low $|\widehat{J}|$ and $\widehat{R}$ are preferred from the point of view of statistical modeling.

Finally, we provide the rates of non-included true variables (denoted by M for Misses), and the rates of incorrectly included variables (FA for False Alarms). Ideally, both rates are low, especially the M-rates, since we do not wish to discard relevant features.

<table>
<thead>
<tr>
<th>Table 2: Performance comparisons between GLASSO, RSC, method 1, method 2 and method 3 in the $p &gt; m$ experiment with $b = 0.5, 1$, $</th>
<th>J</th>
<th>= 15$ and $r = 2$. The measures are 40% trimmed mean of mean squared errors (MSE), median # of predictors ($</th>
<th>\widehat{J}</th>
<th>$), median rank estimate ($\widehat{R}$), and misses (M) and false alarms (FA) rates in variable selection.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b = 0.5$</td>
<td>MSE</td>
<td>$</td>
<td>\widehat{J}</td>
<td>$</td>
</tr>
<tr>
<td>GLASSO</td>
<td>206</td>
<td>10</td>
<td>10</td>
<td>53%</td>
</tr>
<tr>
<td>RSC</td>
<td>485</td>
<td>100</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>method 1</td>
<td>138</td>
<td>19</td>
<td>2</td>
<td>36%</td>
</tr>
<tr>
<td>method 2</td>
<td>197</td>
<td>21</td>
<td>2</td>
<td>33%</td>
</tr>
<tr>
<td>method 3</td>
<td>169</td>
<td>10</td>
<td>2</td>
<td>53%</td>
</tr>
<tr>
<td>$b = 1$</td>
<td>MSE</td>
<td>$</td>
<td>\widehat{J}</td>
<td>$</td>
</tr>
<tr>
<td>GLASSO</td>
<td>511</td>
<td>14</td>
<td>10</td>
<td>41%</td>
</tr>
<tr>
<td>RSC</td>
<td>1905</td>
<td>100</td>
<td>2</td>
<td>0%</td>
</tr>
<tr>
<td>method 1</td>
<td>363</td>
<td>21</td>
<td>2</td>
<td>31%</td>
</tr>
<tr>
<td>method 2</td>
<td>434</td>
<td>25</td>
<td>2</td>
<td>27%</td>
</tr>
<tr>
<td>method 3</td>
<td>402</td>
<td>14</td>
<td>2</td>
<td>41%</td>
</tr>
</tbody>
</table>

We can draw the following conclusions from Tables 2 and 3.

- We see that straightforward variable selection via GLASSO often severely misses some true features in the $p > m$ setup as seen from its high M numbers. RSC
Table 3: Performance comparisons between GLASSO, RSC, method 1, method 2 and method 3 in the $m > p$ experiment with $b = 0.2, 0.4$, $|J| = 15$ and $r = 5$. The measures are 40% trimmed mean of mean squared errors (MSE), median # of predictors ($|\hat{J}|$), median rank estimate ($\hat{R}$), and misses (M) and false alarms (FA) rates in variable selection.

|       | MSE  | $|\hat{J}|$ | $\hat{R}$ | M  | FA  |
|-------|------|-----------|-----------|----|-----|
| $\text{b = 0.2}$ |      |           |           |    |     |
| GLASSO | 18.1 | 14        | 14        | 4% | 1%  |
| RSC    | 11.9 | 25        | 5         | 0% | 100%|
| method 1 | 8.3  | 15        | 5         | 0% | 1%  |
| method 2 | 8.4  | 15.5      | 5         | 0% | 8%  |
| method 3 | 8.9  | 14        | 5         | 4% | 1%  |
| $\text{b = 0.4}$ |      |           |           |    |     |
| GLASSO | 17.7 | 15        | 15        | 0% | 0%  |
| RSC    | 11.5 | 25        | 5         | 0% | 100%|
| method 1 | 8.1  | 15        | 5         | 0% | 0%  |
| method 2 | 8.1  | 15        | 5         | 0% | 7%  |
| method 3 | 8.1  | 15        | 5         | 0% | 0%  |

achieved good rank recovery, as expected, but, by the definition of this estimator, it uses all $p$ variables. Clearly both GLASSO and RSC alone are inferior to the three JRRS-type methods (methods 1, 2 and 3).

- **Method 1** (RSC→RCGL) dominates all other methods. Its MSE results are impressive and confirm the rate improvement established in Section 2.2 over the GLASSO and RSC. While **method 1** may not give exactly $|\hat{J}| = |J| = 15$, its M numbers indicate that we did not miss many true features.

- **Method 2**, unlike **method 1** and **method 3**, did not use the large validation data for ideal parameter tuning, which explains its slight inferiority relative to the
other two methods. However, we see that even without validation-based tuning, which may at times be infeasible in practice, this method is a serious contender. It supports the theoretical findings of Theorem 2 on the usage of the penalty (2) for model comparison and tuning parameter selection.

- The performance of method 3 is inferior to that of method 1 in the $p > m$ experiment, and comparable with both method 1 and method 2 in the $m > p$ experiment, when three methods have essentially the same behavior.

In conclusion, we found that method 1 is the clear winner in terms of performance as well as in computational speed, among the two-stage JRRS procedures we considered, and is particularly appealing in the $m < p$ regime. In particular, it shows the advantage of the novel penalty type which enforces simultaneous (row) sparsity and rank reduction on the coefficient matrix. Method 2 using penalty (2) provides evidence of success of Theorem 2 in dimension reduction.

4 Applications

In this section, we apply method 1 (RSC→RCGL) with the tuning parameters chosen via cross-validation, to two real datasets from machine learning and cognitive neuroscience.

Norwegian Paper Quality These data were obtained from a controlled experiment that was carried out at a paper factory in Norway (Norske Skog, the world’s second-largest producer of publication paper) to uncover the effect of three control variables
$X_1, X_2, X_3$ on the quality of the paper which was measured by 13 response variables. Each of the control variables $X_i$ takes values in $\{-1, 0, 1\}$. To account for possible interactions and non-linear effects second order terms were added to the set of predictors, yielding $X_1, X_2, X_3, X_1^2, X_2^2, X_3^2, X_1 \cdot X_2, X_1 \cdot X_3, X_2 \cdot X_3$ and the intercept term. There were 29 observations with no missing values made on all response and predictor variables. The Box-Behnken design of the experiment and the resulting data are described in Aldrin (1996) and Izenman (2008). Since neither the group penalty nor the rank constraint is imposed on the intercept term, we always center the responses and standardize the predictors in the training data (and transform the validation/test data accordingly).

The data set can be downloaded from the website of Izenman (2008) and its structure clearly indicates that dimension reduction is possible, making it a typical application for reduced rank regression methods. The RSC method with adaptive tuning, as described in Bunea, She, and Wegkamp (2011b), selected the rank $\hat{r} = 3$. This finding is consistent with Aldrin (1996), who assessed the performance of the rank 3 estimator by leave-one-out cross-validation (LOOCV) and obtained a minimum LOOCV error (total squared error, unscaled) of 326.2. We then employed the newly developed method 1 to automatically determine the useful predictors and pursue the optimal projections. Not surprisingly, the selected rank is still 3, yielding 3 new scores, which are now constructed from only 6 of the original 9 predictors, with $X_1^2, X_1 \cdot X_2$ and $X_2 \cdot X_3$ discarded, and only the variables from Table 4 selected. The tuning result was the same for 10-fold CV and LOOCV. The minimum LOOCV error is now 304.5. We found no interaction effect between $X_2$ and $(X_1, X_3)$, an interesting complement to Aldrin’s analysis.
Table 4 shows the construction weights of the 3 new orthogonal score variables from
the rank-3 RSC on the selected set of variables. They are ordered by an importance
measure given by the associated eigenvalues of $Y'X(X'X)^{-1}X'Y =: W$ (see Reinsel
and Velu (1998) and Izenman (2008) for the explanation). For instance, the first im-
portant score variable (accounting for 57.5% of the trace of $W$) can be roughly read as
$2X_1 - 0.5X_2 - X_3 - 0.5X_2^2 + X_3^2 + X_1X_3$, or simply $2X_1 + X_1X_3 + 1_{x_3=-1} \cdot 1_{x_2=1}$. This
can be used as a concise summary predictor for all 13 response variables simultaneously
and it quantifies the effect of the design variables on paper quality control.

Table 4: Paper quality control: Joint new feature construction from the selected pre-
dictors with $r = 3, |\hat{J}| = 6$. The extracted components are ordered by their associated
eigenvalues.

<table>
<thead>
<tr>
<th>New Scores</th>
<th>Eigenvalues</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_2^2$</th>
<th>$X_3^2$</th>
<th>$X_1 \cdot X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>112.4</td>
<td>1.9244</td>
<td>-0.5288</td>
<td>-1.2321</td>
<td>-0.4443</td>
<td>1.3109</td>
<td>1.1898</td>
</tr>
<tr>
<td>2</td>
<td>40.7</td>
<td>0.8231</td>
<td>-0.5937</td>
<td>0.9324</td>
<td>0.7819</td>
<td>-0.1599</td>
<td>-0.8536</td>
</tr>
<tr>
<td>3</td>
<td>24.9</td>
<td>0.2871</td>
<td>1.0336</td>
<td>0.4215</td>
<td>0.8365</td>
<td>0.4245</td>
<td>0.4677</td>
</tr>
</tbody>
</table>

Cognitive Neuroimaging  We present an analysis of the data set described in Bunea
et al. (2011a) and collected to investigate the effect of the HIV-infection on human
cognitive abilities. Neuro-cognitive performance is typically measured via correlated
neuro-cognitive indices (NCIs). This study employed $n = 13$ NCIs, falling into five
domains of attention/working memory, speed of information processing, psychomotor
abilities, executive function, and learning and memory. These indices were measured
for 62 HIV+ patients in the study. The set of explanatory variables was large and
contained: (a) clinical and demographic predictors and (b) brain volumetric and diffu-
sion tensor imaging (DTI) derived measures of several white-matter regions of interest, such as fractional anisotropy, mean diffusivity, axial diffusivity, and radial diffusivity, along with all volumetrics × DTI interactions. We refer to Bunea et al. (2011a) for details. The final model has $p = 235$ predictors, much greater than the sample size $m = 62$. An initial analysis of this data set was performed using the RSC to select a model of rank 1 and construct the corresponding new predictive score. Although this is a massive reduction of the dimension of the predictor space, all 235 initial predictors were involved in the construction of the new score – see the left panel of Figure 1. This leaves unanswered the important question as to what variables (especially which DTI derived measures) are most predictive of the neuro-cognitive changes in HIV+ patients. After standardizing the predictors, we run method 1. We selected a model of rank 1 and constructed one new predictive score but, very importantly, this score is a linear combination of only 10 predictors that were selected from the original pool of 235. See Table 5 for the description of these predictors and Figure 1 for the display of the weights vector used to construct the new score variable via RSC and RCGL, respectively. When we set aside 30% of the data as a separate test set, and used the remaining to fit the model and tune the regularization parameters, the mean squared error (MSE) of the RSC estimate was 192.9, while the MSE of the newly proposed method was only 138.4. Moreover, our analysis demonstrates not only the existence of a strong association between Education and the neuro-cognitive abilities of HIV+ patients, which had already been established by other means in the literature, but also suggests, as a perhaps new finding, that the fractional anisotropy at corpus callosum (fa_cc1) stands out among the very many DTI-derived measures, in terms of predictive
Table 5: Cognitive Neuroimaging: Selected cognitive variables for constructing the new score variable.

<table>
<thead>
<tr>
<th>Selected measures</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>duration of education in years</td>
</tr>
<tr>
<td>kmsk_cocopi</td>
<td>with cocaine history or not</td>
</tr>
<tr>
<td>HIV_stage*whitematter</td>
<td>(Nadir CD4 (cells/µl) &gt; 200 or not) × (size of white matter)</td>
</tr>
<tr>
<td>fa_cc1*fa_cc1</td>
<td>fractional anisotropy measured at corpus callosum subregion 1, squared</td>
</tr>
<tr>
<td>fa_cc1*whitematter</td>
<td>(fractional anisotropy at corpus callosum subregion 1) × (size of white matter (left and right))</td>
</tr>
<tr>
<td>fa_cc1*thalamus</td>
<td>(fractional anisotropy at corpus callosum subregion 1) × (volumetric measure of thalamus)</td>
</tr>
<tr>
<td>fa_cc1*putamen</td>
<td>(fractional anisotropy at corpus callosum subregion 1) × (volumetric measure of putamen)</td>
</tr>
<tr>
<td>fa_cc1*cc_body</td>
<td>(fractional anisotropy at corpus callosum subregion 1) × (volumetric measure of corpus callosum at body)</td>
</tr>
<tr>
<td>md_cc5*caudate</td>
<td>(mean diffusivity at corpus callosum subregion 5) × (volumetric measure of caudate)</td>
</tr>
<tr>
<td>md_ic3*putamen</td>
<td>(mean diffusivity at internal capsule subregion 3) × (volumetric measure of putamen)</td>
</tr>
</tbody>
</table>

Appendix

Proof of Theorem 1

Proof. By definition, for any $p \times n$ matrix $B$, the inequality

$$\|Y - X\tilde{B}\|^2 + pen(\tilde{B}) \leq \|Y - XB\|^2 + pen(B)$$
Next we note that, writing the right column coincides with

\[ \sum_{i} \mathbf{x}_i = \mathbf{0} \]

holds, and this is equivalent with

\[
\| X A - X \hat{B} \|^2 \leq \| X A - X B \|^2 + 2 \text{pen}(B) \\
+ \left\{ 2 E, X \hat{B} - XB > -\text{pen}(\hat{B}) - \text{pen}(B) \right\}. 
\] (15)

Next we note that, writing \( J = J(B) \) and \( \hat{J} = J(\hat{B}) \),

\[
< E, X(\hat{B} - B) > = < E, X_{J \cap \hat{J}}(\hat{B} - B) > + < E, X_{J \cap \hat{J}}(\hat{B} - B) > \\
+ < E, X_{J \cap \hat{J}}(\hat{B} - B) > + < E, X_{J \cap \hat{J}}(\hat{B} - B) >,
\]

where \( X_I \), for a generic index set \( I \), is defined as follows: its \( i \)th column coincides with that of \( X \) if \( i \in I \), otherwise we set the entire column to zero. Furthermore, writing
\( r = r(B) \) and \( \hat{r} = r(\hat{B}) \), we have

\[
|J| r + |\tilde{J}| \hat{r} = |J \cap \tilde{J}| (r + \hat{r}) + |J \cap \tilde{J}^c| r + |J^c \cap \tilde{J}| \hat{r}.
\]

Combining the last two displays and using the penalty term given by (2), the last term in (15) above becomes

\[
\begin{align*}
\left\{ 2 < E, X \hat{B} - XB > -\text{pen}(\hat{B}) - \text{pen}(B) \right\} \\
= 2 < E, X_{J \cap \tilde{J}}(\hat{B} - B) > -c\sigma^2 (n + |J \cap \tilde{J}|) (r + \hat{r}) \\
+ 2 < E, X_{J \cap \tilde{J}^c}(\hat{B} - B) > -c\sigma^2 (n + |J \cap \tilde{J}^c|) r \\
+ 2 < E, X_{J^c \cap \tilde{J}}(\hat{B} - B) > -c\sigma^2 (n + |J^c \cap \tilde{J}|) \hat{r}.
\end{align*}
\]

Define \( P_I \) to be the projection matrix onto the column space of \( X_I \), i.e., \( P_I = X_I (X_I' X_I)^{-1} X_I' \).

Recall that for any matrices \( B \) and \( C \) we have \( < B, C > \leq \rho(B) \| C \|_F \), where \( \rho(B) \) denotes the largest singular value of \( B \). Using this inequality and the fact that \( 2xy \leq a^{-1} x^2 + ay^2 \), for any \( a > 0 \), we further bound the terms in the right hand side of the display above by

\[
\begin{align*}
& a^{-1} d_1^2 (P_{J \cap \tilde{J}} E)(\hat{r} + r) + a\|X_{J \cap \tilde{J}}(\hat{B} - B)\|^2 - c\sigma^2 (n + |J \cap \tilde{J}|)(\hat{r} + r) \\
+ & a^{-1} d_1^2 (P_{J \cap \tilde{J}^c} E)r + a\|X_{J \cap \tilde{J}^c}(\hat{B} - B)\|^2 - c\sigma^2 (n + |J \cap \tilde{J}^c|) r \\
+ & a^{-1} d_1^2 (P_{J^c \cap \tilde{J}} E)\hat{r} + a\|X_{J^c \cap \tilde{J}}(\hat{B} - B)\|^2 - c\sigma^2 (n + |J^c \cap \tilde{J}|) \hat{r},
\end{align*}
\]
and this can be further bounded by

\[ a\|X(\hat{B} - B)\|^2 + 3n \max_{I} (a^{-1}d^2_I(P_I E) - c\sigma^2(n + |I|))_+ , \]

and so

\[
\left\{ 2 < E, X\hat{B} - XB > -\text{pen}(\hat{B}) - \text{pen}(B) \right\} \\
\leq a\|X(\hat{B} - B)\|^2 + 3n \max_{I} (a^{-1}d^2_I(P_I E) - c\sigma^2(n + |I|))_+ .
\]

Next, since \(a\|X(\hat{B} - B)\|^2 \leq a(1 + b)\|X\hat{B} -XA\|_{\hat{F}}^2 + a(1 + 1/b)\|XB -XA\|_{\hat{F}}^2 \) for any \( b > 0 \), inequality (15) together with the display above used with \( c = a^{-1}(1 + t)^2 \), for any given \( t > 0 \), and some \( a < 1 \) such that \( a(1 + b) < 1 \) and \( c > 1 \), yields

\[
(1 - a - ab)\mathbb{E} \left[ \|X\hat{B} - XA\|_{\hat{F}}^2 \right] \leq (1 + a + a/b)\|XB -XA\|_{\hat{F}}^2 + 2\text{pen}(B) \\
+ 3na^{-1}\mathbb{E} \left[ \max_{I} (d^2_I(P_I E) - (1 + t)^2\sigma^2(n + |I|))_+ \right],
\]

which, by Lemma 9 below further yields, for any matrix \( B \),

\[
(1 - a - ab)\mathbb{E} \left[ \|X\hat{B} - XA\|_{\hat{F}}^2 \right] \\
\leq (1 + a + a/b)\|XB -XA\|_{\hat{F}}^2 + 2\text{pen}(B) + \frac{3n 2\sigma^2}{a} \frac{1 + t}{t} \frac{1}{1 - \exp(-t^2/2)} \exp(-nt^2/2) \\
= (1 + a + a/b)\|XB -XA\|_{\hat{F}}^2 + \left\{ 2 + \frac{3\exp(-nt^2/2)}{(1 + t)(1 - \exp(-t^2/2))} \right\} \text{pen}(B) \\
\equiv (1 + a + a/b)\|XB -XA\|_{\hat{F}}^2 + C_t\text{pen}(B)
\]

since, for this choice of \( c \), \( \text{pen}(B) \) equals \( a^{-1}(1 + t)^2\sigma^2(2n + |J(B)|)r(B) \). Dividing by
1 − a − ab > 0 completes the proof. □

Lemma 9. For any \( t > 0 \) we have

\[
\mathbb{E} \left[ \max_J \left( d_1^2 (P_J E) - (1 + t)^2 \sigma^2 (n + |J|) \sigma^2 \right) \right] \leq \frac{2\sigma^2}{t} \frac{(1 + t)}{1 - \exp(-t^2/2)} \exp(-nt^2/2).
\]

Moreover, for any index set \( J, \) possibly random and possibly dependent on the error matrix \( E \) we have

\[
\mathbb{E} \left[ d_1^2 (P_J E) \right] \leq C_0 \sigma^2 (n + \mathbb{E}[|J|]), \tag{16}
\]

for a universal numerical constant \( C_0 \leq 7.82. \)

Proof. First notice that \( d_1^2 (P_J E) \) depends on \( J \) through the cardinality of \( J \) only. Indeed, \( d_1^2 (P_J E) \) has the same distribution as \( d_1^2 (Z_k) \) with \( k = |J| \) and \( Z_k \) a \( m \times k \) matrix of independent \( N(0, \sigma^2) \) entries. Consequently

\[
\begin{align*}
\mathbb{E} & \left[ \max_J \left( d_1^2 (P_J E) - (1 + t)^2 (n + |J|) \sigma^2 \right) \right] \\
& \leq \mathbb{E} \left[ \max_k \left( d_1^2 (Z_k) - (1 + t)^2 (n + k) \sigma^2 \right) \right] \\
& \leq \sum_{k=0}^q \mathbb{E} \left[ (d_1^2 (Z_k) - (1 + t)^2 (n + k) \sigma^2) \right].
\end{align*}
\]

By Lemma 16, page 1307 in Bunea, She, and Wegkamp (2011b), we have, for each \( k, \)

\[
\mathbb{E} \left[ (d_1^2 (Z_k) - (1 + t)^2 (n + k) \sigma^2) \right] \leq 2\sigma^2 (1 + t^{-1}) \exp(-t^2 (n + k)/2), \quad \text{and so}
\]
\[ \mathbb{E} \left[ \max_j (d_1^2(P_J E) - (1 + t)^2(n + |J|)\sigma^2)_{+} \right] \leq \frac{2\sigma^2}{t} \frac{(1 + t)}{1 - \exp(-t^2/2)} \exp(-nt^2/2), \]

which concludes the first part of our proof. The second part follows immediately since we also have

\[ \mathbb{E} \left[ \max_j (d_1^2(P_J E) - (1 + t)^2(n + |J|)\sigma^2) \right] \leq \frac{2\sigma^2}{t} \frac{(1 + t)}{1 - \exp(-t^2/2)} \exp(-nt^2/2), \]

and thus there exists a constant \( C_0 \) independent of \( J \) and the design for which the conclusion holds, and the value of \( C_0 \) is obtained by minimizing the convex function \((1 + t)^2 + 2(1 + 1/t)\exp(-t^2/2)/(1 - \exp(-t^2/2))\). \( \square \)

### 4.1 Proof of Theorem 2

**Proof.** By the same reasoning as in the proof of Theorem 1, we obtain

\[ (1 - a - ab) \left[ \|X\tilde{B} - XA\|_F^2 \right] \leq (1 + a + a/b)\|XB_j - XA\|_F^2 + 2\text{pen}(B_j) + 3n \left[ \max_j (a^{-1}d_1^2(P_J E) - c\sigma^2(n + |I|)_{+}) \right] \]

for any \( B_j \), random or not. Taking expectation on both sides of the inequality and applying Lemma 9 gives the result. \( \square \)
4.2 Proof of Theorem 3

Proof. By the definition of $\tilde{B}_k$ we get

$$\|X\tilde{B}_k - XA\|_F^2 \leq \|XB - XA\|_F^2 + 2 < E, X(\tilde{B}_k - B) > + 2\lambda \|B\|_{2,1} - \|\tilde{B}_k\|_{2,1}$$

We denote the rows of $\tilde{B}_k$ by $\tilde{b}_1, \ldots, \tilde{b}_p$, and let $\tilde{J}_k = \{i : \tilde{b}_i \not= 0\}$ the index set of non-zero rows. Let $B$ be any matrix with rows $b_1, \ldots, b_p \in \mathbb{R}^n$ with $r(B) \leq k$ and $J = \{j : b_j \not= 0\}$ such that $\Sigma = X'X/m$ satisfies condition $\mathfrak{A}(J, \delta_J)$. The notation $X_J$ is as before in the proof of Theorem 1 and we use $\tilde{X}$ to denote $X_{\tilde{J}}$ with $\tilde{J} = \tilde{J}_k \cup J$. Finally, we write $\Delta^2 = \|XB - XA\|_F^2$ and $\tilde{\Delta}^2 = \|X\tilde{B}_k - XA\|_F^2$. In this new notation, we have

$$\tilde{\Delta}^2 \leq \Delta^2 + 2 < E, \tilde{X}(\tilde{B}_k - B) > + 2\lambda \|B\|_{2,1} - \|\tilde{B}_k\|_{2,1}$$

$$\leq \Delta^2 + 2 < E, \tilde{X}(\tilde{B}_k - B) > + 2\lambda \|B - \tilde{B}_k\|_{2,1}.$$  \hspace{1cm} (17)

The second term on the right of (17) is handled as follows. For any $\alpha > 0$

$$2 < E, \tilde{X}(\tilde{B}_k - B) > = 2 < P_J E, \tilde{X}(\tilde{B}_k - B) >$$

$$\leq 2d_1(P_J E) \|\tilde{X}(\tilde{B}_k - B)\|_1$$

$$\leq 2d_1(P_J E) \sqrt{2k} \|\tilde{X}(\tilde{B}_k - B)\|_F$$

$$\leq 2d_1(P_J E) \sqrt{2k} \|\tilde{X}(\tilde{B}_k - A)\|_F + 2d_1(P_J E) \sqrt{2k} \|\tilde{X}(B - A)\|_F$$

$$\leq 4\alpha k d_1^2(P_J E) + \frac{\tilde{\Delta}^2 + \Delta^2}{\alpha}$$
using \( r(\hat{B} - B) \leq r(B_k) + r(B) \leq 2k \) and the inequality \( 2xy \leq \alpha x^2 + y^2/\alpha \) for all \( \alpha > 0 \).

Let \( a_1, \ldots, a_p \in \mathbb{R}^n \) denote the rows and of the matrix \( A \). Then, for the last term in (17), we have, for any \( \beta > 0 \),

\[
2\lambda \| (B - \hat{B}_k)_J \|_{2,1} = 2\lambda \sum_{i \in J} \| b_i - \hat{b}_i \|_2 \\
\leq 2\lambda \sum_{i \in J} (\| b_i - a_i \|_2 + \| a_i - \hat{b}_i \|_2) \\
\leq \frac{2\beta}{m} \lambda^2 |J| + \frac{m}{\beta} \sum_{i \in J} (\| b_i - a_i \|_2^2 + \| a_i - \hat{b}_i \|_2^2) \\
\leq \frac{2\beta}{m} \lambda^2 |J| + \frac{1}{\beta \delta_J} (\Delta^2 + \Delta^2).
\]

The last inequality uses our assumption on \( \Sigma \). Indeed, \( \Sigma \) satisfies Assumption \( \mathfrak{A}(J, \delta_J) \) implies that for all \( p \times n \) matrices \( M \) with rows \( m_i \) and columns \( \tilde{m}_j = (\tilde{m}_{j,1}, \ldots, \tilde{m}_{j,p}) \), we have

\[
\text{tr}(M^T \Sigma M) = \sum_j \text{tr}(\tilde{m}_j^T \Sigma \tilde{m}_j) \geq \sum_j \delta_I \sum_{k \in I} \| \tilde{m}_{j,k} \|_2^2 = \delta_I \sum_{i \in I} \| m_i \|_2^2.
\]

By Lemma 9 we have

\[
\mathbb{E} \left[ d_1^2(P_J E) \right] \leq C_0 \left( n + \mathbb{E} \left[ |\hat{J}| \right] \right) \sigma^2
\] (18)
for some constant $C_0$ large enough, see (16). So far we proved that

$$
\mathbb{E} \left[ \hat{\Delta}^2 \right] \leq \left( 1 + \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \Delta^2 + \left( \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \mathbb{E} \left[ \hat{\Delta}^2 \right] + 4\alpha k \mathbb{E} \left[ d_J^2 (P_J E) \right] + 2\beta \frac{\lambda^2}{m} |J|
$$

$$
\leq \left( 1 + \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \Delta^2 + \left( \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \mathbb{E} \left[ \hat{\Delta}^2 \right] + 4\alpha k \sigma^2 C_0 (n + |J|) + \mathbb{E} \left[ |J_k \setminus J| \right] + 2\beta \frac{\lambda^2}{m} |J|.
$$

Taking $\lambda^2 = 2LC_0 \rho m k \sigma^2$ as in (19) and using inequality (20) of Lemma 10 below, with $L$ suitably large, we obtain

$$
\mathbb{E} \left[ \hat{\Delta}^2 \right] \leq \left( 1 + \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \Delta^2 + \left( \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \mathbb{E} \left[ \hat{\Delta}^2 \right] + 4\alpha k C_0 \sigma^2 (n + |J|) + 4\alpha \sigma^2 \mathbb{E} \left[ |J_k| \right] + 2\beta \frac{\lambda^2}{m} |J|
$$

$$
\leq \left( 1 + \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \Delta^2 + \left( \frac{1}{\alpha} + \frac{1}{\beta \delta_J} + \frac{4\alpha}{L - 1} \right) \mathbb{E} \left[ \hat{\Delta}^2 \right] + 4\alpha k C_0 \sigma^2 \left( n + |J| \right) + 4\beta \lambda^2 \frac{2LC_0 \rho m k \sigma^2}{m} |J|
$$

$$
= \left( 1 + \frac{1}{\alpha} + \frac{1}{\beta \delta_J} \right) \Delta^2 + \left( \frac{1}{\alpha} + \frac{1}{\beta \delta_J} + \frac{4\alpha}{L - 1} \right) \mathbb{E} \left[ \hat{\Delta}^2 \right] + 4\alpha C_0 \sigma^2 \left( n + |J| \right) k + 4\beta C_0 L \rho |J| k \sigma^2 + 4 \frac{\alpha C_0}{L - 1} nk \sigma^2.
$$

This yields for choosing $L, \alpha$ and $\beta \sim 1/\delta_J$ large enough,

$$
\mathbb{E} \left[ \hat{\Delta}^2 \right] \lesssim \Delta^2 + \frac{\rho}{\delta_J} k |J| \sigma^2 + (n + |J|) k \sigma^2.
$$

This concludes our proof of risk bound (8).  \qed
Lemma 10. Let $\hat{B}_k$ be a minimizer of (12) with tuning parameter

$$\lambda^2 = 2LC_0\rho mk\sigma^2$$

with $\rho = \lambda_1(X'X)/m = \lambda_1(\Sigma)$ and for some constant $C_0 < 7.82$ as in (16). Then, for any $L > 1$,

$$\mathbb{E}\left[|J(\hat{B}_k)|\right] \leq \frac{1}{(L-1)C_0k\sigma^2} \mathbb{E}\left[\|X\hat{B}_k - XA\|^2_F\right] + \frac{n}{L-1}. \quad (20)$$

Proof. Let $J = J(\hat{B})$. Since $r(\hat{B}) \leq k$, we write $\hat{B}_k = \hat{S}\hat{V}_k'$ for some $\hat{V}_k$ with $k$ columns satisfying $\hat{V}_k'\hat{V}_k = I_{k \times k}$, . Following the lines of argument in the proof of Lemma 11, with $V$ fixed at $\hat{V}_k$, $\hat{S}$ is the (globally) optimal solution to the convex problem of (25).

Let $X = [x_1 \cdots x_p]$. Using the KKT equation of $\hat{S}' = [\hat{s}_1 \cdots \hat{s}_p]$, we obtain

$$\|\hat{b}_i\| \neq 0 \Rightarrow \|\hat{s}_i\| \neq 0 \Rightarrow \|x_i'(X\hat{S} - Y\hat{V}_k)\|_2 = \lambda,$$  \quad (21)

from which it follows that

$$|\hat{J}_k|\lambda^2 = \sum_{j \in \hat{J}_k} \|x_j'(X\hat{S} - XA\hat{V}_k - E\hat{V}_k)\|_2^2$$

$$= \sum_{j \in \hat{J}_k} \|x_j'(X\hat{S} - XA\hat{V}_k - P_{\hat{J}_k}E\hat{V}_k)\|_2^2$$

$$\leq 2\sum_{j \in \hat{J}_k} \|x_j'(X\hat{S} - XA\hat{V}_k)\|_2^2 + 2\sum_{j \in \hat{J}_k} \|x_j'P_{\hat{J}_k}E\hat{V}_k\|_2^2$$

$$\leq 2\lambda_1(X'X)\|X\hat{B}_k - XA\|^2_F + 2\lambda_1(X'X)\|P_{\hat{J}_k}E\hat{V}_k\|^2_F$$

$$\leq 2\lambda_1(X'X)\left[\tilde{\Delta}^2 + kd_1^2(P_{\hat{J}_k}E)\right]$$

42
since \( r(\hat{V}_k) \leq k \).

Reasoning as in (18), for the same constant \( C_0 \),

\[
\mathbb{E} \left[ d_r^2(P_{\hat{J}_k}E) \right] \leq C_0\sigma^2\{n + \mathbb{E}[|\hat{J}_k|]\}
\]

(22)

so that

\[
\mathbb{E} \left[ |\hat{J}_k| \right] \leq \frac{2\lambda_1(X'X)}{\lambda^2} \left( \mathbb{E} \left[ \hat{\Delta}^2 \right] + C_0\sigma^2nk + kC_0\sigma^2\mathbb{E} \left[ |\hat{J}_k| \right] \right)
\]

(23)

Choosing \( \lambda^2 \) as in (19), we immediately obtain (20).

\[ \Box \]

### 4.3 Proof of Theorem 5.

**Proof.** Recall that \( \hat{r} \) is the number of eigen-values of \( Y'PY \) that exceed the threshold level \( 2\mu = 4\sigma^2(\sqrt{n} + \sqrt{q})^2 \). Theorem 2 and Corollary 4 in Bunea, She, and Wegkamp (2011b) show that for \( c_0 = (\sqrt{2} - 1)^2/2 \),

\[
\mathbb{P}\{\hat{r} \neq r\} \leq \exp\{-c_0(n + q)\}.
\]

Next, we decompose the risk as follows:

\[
\mathbb{E} \left[ \|X_A - X\hat{B}^{(1)}\|_F^2 \right] = \mathbb{E} \left[ \|X_A - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} = r\} \right] + \mathbb{E} \left[ \|X_A - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \neq r\} \right]
\]

(24)
The first term on the right gives the bound obtained in Theorem 3 for \( k = r \). It remains to bound the second term. Let \( O \) denote the \( p \times n \) matrix with all entries equal to zero. Then, since \( \hat{r} = r(\hat{B}) \geq r(O) = 0 \), we have the inequality

\[
\|Y - X\hat{B}^{(1)}\|_F^2 + 2\lambda\|\hat{B}^{(1)}\|_{2,1} \leq \|Y - XO\|_F^2 + 2\lambda\|O\|_{2,1} = \|Y\|_F^2
\]

by the minimizing property of \( \hat{B}^{(1)} \). Let \( P \) be the projection matrix onto \( X \). Using Pythagoras, it follows that \( \|PY - X\hat{B}^{(1)}\|_F^2 \leq \|PY\|_F^2 \). Consequently,

\[
\begin{align*}
\mathbb{E} \left[ \|XA - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \neq r\} \right] & \leq 2\mathbb{E} \left[ \left\{ \|PY - XA\|_F^2 + \|PY - X\hat{B}^{(1)}\|_F^2 \right\} I\{\hat{r} \neq r\} \right] \\
& \leq 2\mathbb{E} \left[ \left\{ \|PE\|_F^2 + \|PY\|_F^2 \right\} I\{\hat{r} \neq r\} \right] \\
& \leq 2\mathbb{E} \left[ \left\{ 3\|PE\|_F^2 + 2\|XA\|_F^2 \right\} I\{\hat{r} \neq r\} \right] \\
& = 2\mathbb{E} \left[ 3\|PE\|_F^2 I\{\hat{r} \neq r\} \right] + 4\|XA\|_F^2 \mathbb{P}\{\hat{r} \neq r\}.
\end{align*}
\]

Since \( \|PE\|_F^2 / \sigma^2 \) has a Chi-square distribution with \( qn \) degrees of freedom (cf Lemma 3 in Bunea, She, and Wegkamp (2011b)),

\[
\mathbb{E} \left[ \|PE\|_F^4 \right] \leq (q^2 n^2 + 2qn)\sigma^4 \leq 2q^2 n^2 \sigma^4.
\]

We obtain, using the Cauchy-Schwarz inequality,

\[
\mathbb{E} \left[ \|PE\|_F^2 I\{\hat{r} \neq r\} \right] \leq \sqrt{2}qn\sigma^2 \exp\{-c_0(n + q)/2\},
\]

44
and so

\[ \mathbb{E}\left[\|XA - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \neq r\}\right] \leq 4\|XA\|_F^2 \exp\{-c_0(n + q)\} + \sqrt{2qn}\sigma^2 \exp\{-c_0(n + q)/2\}. \]

The second term on the right is clearly of order $O(1)$. For the first term on the right, we invoke condition $C_2$ on $\|XA\|_F^2$, and we see that $\mathbb{E}\left[\|XA - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \neq r\}\right]$ is of order $(n + |J|)r$. Hence both terms on the right hand side of in (24) are of order $(n + |J|)r$. This completes our proof. \qed

### 4.4 Proof of Theorem 6.

**Proof.** From Theorem 2 of Section 2, applied to the random matrices $\hat{B}_{\lambda,k}$, we immediately obtain the oracle

\[ \mathbb{E}[\|X\hat{B}^{(2)} -XA\|^2] \lesssim \min_{k,\lambda} \mathbb{E}\left[\|X\hat{B}_{\lambda,k} -XA\|^2 + (n + |J(\hat{B}_{\lambda,k})|)k\right]. \]

This result shows that $\hat{B}^{(2)}$ encourages sparse row and rank approximations. Since $\Sigma$ satisfies $\mathcal{A}(J(A), \delta_{J(A)})$, Theorem 3 yields, for each global solution $\hat{B}_{\lambda,k}$,

\[ \mathbb{E}[\|XA - X\hat{B}_{\lambda,k}\|^2] \lesssim \left\{\frac{n + |J(A)|}{m\delta_{J(A)}}\right\} k\sigma^2, \]
for \( \lambda = \lambda(C) \) of the form (7) with \( C \) large enough. The two previous inequalities combined with (10) and Lemma 10 yield the oracle bound

\[
\mathbb{E}[\|X\hat{B}^{(2)} - XA\|^2] \lesssim \min_{k,\lambda} \mathbb{E} \left[ \|X\hat{B}_{\lambda,k} - XA\|^2 + (n + |J(\hat{B}_{\lambda,k})|)r(\hat{B}_{\lambda,k}) \right] \\
\lesssim \left( \frac{\lambda_1(X'X)}{m\delta_{J(A)}} |J(A)| + n \right) r\sigma^2.
\]

Since Theorem 6 assumes that \( \lambda_1(X'X)/\{m\delta_{J(A)}\} = O(1) \), the claim follows. \( \square \)

4.5 Proof of Theorem 7

In this proof, we drop the subscripts \( \lambda, k \) in the \((S, V)\) iterates.

Proof of Part (ii) The proof of this part of Theorem 7 will follow from the Global Convergence Theorem (GCT) of Zangwill (1969). For completeness, we state this theorem below, then we verify that its conditions hold in Lemmas 11, 12 and 13.

The Global Convergence Theorem (Luenberger and Ye, 2008). Let \( A \) be a map describing an algorithm on \( X \) and suppose that given \( x_0 \) the sequence \( \{x_k\}_{k=1}^\infty \) is generated by \( x_{k+1} \in A(x_k) \). Let a solution set \( \Gamma \subset X \) be given, and suppose:

1. All points \( x_k \) are contained in a compact set \( S \subset X \);

2. There exists a continuous function \( Z \) on \( X \) such that (a) if \( x \notin \Gamma \), then \( Z(y) < \)
   \( Z(x) \) for all \( y \in A(x) \) (b) if \( x \in \Gamma \), then \( Z(y) \leq Z(x) \) for all \( y \in A(x) \);

3. The mapping \( A \) is closed at points outside \( \Gamma \).
Then the limit of any convergent subsequence of \( \{x_k\}_k \) is a solution.

We begin by introducing a map, usually referred to in the literature as a point-to-set map, to characterize our algorithm \( \mathcal{A} \). Let \( \Omega = \mathbb{R}^{p \times k} \times \mathbb{O}^{n \times k} \). Define \( \mathcal{M}^S : \Omega \to 2^{\Omega}, \mathcal{M}^V : \Omega \to 2^{\Omega} \) as follows

\[
\mathcal{M}^S(S, V) = \{ (\bar{S}, \bar{V}) \in \Omega : \inf_{\tilde{S} \in \mathbb{R}^{p \times k}} F(\tilde{S}, \bar{V}) \geq F(\bar{S}, V) \},
\]

\[
\mathcal{M}^V(S, V) = \{ (S, \bar{V}) \in \Omega : \inf_{\tilde{V} \in \mathbb{O}^{n \times k}} F(S, \tilde{V}) \geq F(S, V) \},
\]

and define \( \mathcal{M} = \mathcal{M}^V \mathcal{M}^S \) as a composite point-to-set map, see Luenberger and Ye (2008) for more details. Algorithm \( \mathcal{A} \) can be described by

\[(S^{(j+1)}, V^{(j+1)}) \in \mathcal{M}(S^{(j)}, V^{(j)}),\]

that is, \((S^{(j+1)}, V^{(j)}) \in \mathcal{M}^S(S^{(j)}, V^{(j)})\) and \((S^{(j+1)}, V^{(j+1)}) \in \mathcal{M}^V(S^{(j+1)}, V^{(j)})\).

Recall that

\[
F(S, V; \lambda) = \frac{1}{2} \| Y - XSV' \|_F^2 + \lambda \| SV' \|_{2,1}, \ S \in \mathbb{R}^{p \times k}, \ V \in \mathbb{O}^{n \times k},
\]

and that we analyze the unconstrained minimum of \( F \) over the product manifold \( \mathbb{R}^{p \times k} \times \mathbb{O}^{n \times k} \). For simplicity, we will write just \( F(B) \) and \( F(S, V) \) for \( F(B; \lambda) \) and \( F(S, V; \lambda) \), respectively, when there is no ambiguity.
Lemma 11 shows the algorithm converges globally for any initial starting point.

**Lemma 11.** For any $j \geq 0$, $\text{rank}(B^{(j)}) \leq k$ and $F(B^{(j)}) \geq F(B^{(j+1)})$.

**Proof.** We write $F(S, V) = \frac{1}{2}\|YV - XS\|_F^2 + \lambda\|S\|_{2,1} + \frac{1}{2}\text{tr}(Y(I - VV')Y')$. Given $V$, (13) reduces to the following optimization problem after vectorization:

$$
\frac{1}{2}\|\text{vec}((YV)') - (X \otimes I)\text{vec}(S')\|_F^2 + \lambda\|S\|_{2,1}, \tag{25}
$$

where ‘vec’ is the standard vectorization operator, and $\otimes$ is the Kronecker product. This is a GLASSO-type optimization problem that is convex in $S$. The global minimum of (25) can always be achieved at some $S$ with $\|S\| < \infty$. Given $S$, writing $F(S, V) = -\text{tr}(Y'XS'V') + \|Y\|_F^2/2 + \|XS\|_F^2/2 + \lambda\|S\|_{2,1}$, we see the optimization problem is equivalent to

$$
\max_{V \in O^{n \times p}} \text{tr}(W'V), \tag{26}
$$

where $W = Y'XS \in \mathbb{R}^{n \times k}$. The (global) maximum can be attained, too, due to the compactness of $O^{n \times p}$. In fact, $\text{tr}(W'V) \leq \sum_i d_i(W)$ by von Neumann’s trace inequality. Let $W = U_wD_wV'_w$ be the SVD with $V_w \in O^{k \times k}$. Then

$$
\hat{V} = U_wV'_w \tag{27}
$$

achieves the upper bound $\sum_i d_i(W)$. This globally optimal solution to (26) is the one
used in Algorithm $\mathcal{A}$. Therefore, we have

$$F(V^{(j)}), S^{(j)}) \geq F(V^{(j)}, S^{(j+1)}) \geq F(V^{(j+1)}, S^{(j+1)}),$$

that is, $F(B^{(j)}_{k,\lambda}) \geq F(B^{(j+1)}_{k,\lambda})$ during each iteration.

**Lemma 12.** Suppose $\lambda > 0$. Then $B^{(j)}$, $S^{(j)}$, $V^{(j)}$ in $\mathcal{A}$ are uniformly bounded in $j$.

**Proof.** From Lemma 11, $\|S^{(j)}\|_2 \leq F(S^{(j)}, V^{(j)}) / \lambda \leq F(S^{(1)}, V^{(1)}) / \lambda \leq F(0, V^{(0)}) / \lambda \leq \|Y\|_F^2 / (2\lambda)$. Therefore, $\|S^{(j)}\|$ must be uniformly bounded.

**Lemma 13.** Suppose $\lambda > 0$. The map $\mathcal{M}$ introduced for describing $\mathcal{A}$ is a closed point-to-set map on $\Omega$.

**Proof.** Notice that

- The set $\Omega = \mathbb{R}^{p \times k} \times \mathbb{O}^{n \times k}$ is closed because $\mathbb{O}^{n \times k}$ is the inverse image of $\{I\}$ under the continuous function $M \mapsto M^T M$, $\forall M \in \mathbb{R}^{n \times k}$; $\mathbb{O}^{n \times k}$ is in fact an embedded submanifold of $\mathbb{R}^{n \times k}$.

- $\mathcal{M}^S(\omega) \neq \emptyset$, $\mathcal{M}^V(\omega) \neq \emptyset$, $\forall \omega \in \Omega$, seen from the proof of Lemma 11.

First, we prove that $\mathcal{M}^S$ is closed on $\Omega$. It suffices to show the point-to-set map $\mathcal{M}^S(x) = \{y \in \mathbb{R}^{p \times k} : F(y, x) \leq \min_{\hat{y} \in \mathbb{R}^{p \times k}} F(\hat{y}, x)\}$ is closed at any $x \in \mathbb{O}^{n \times k}$. Let $x_j \rightarrow x^*$, $y_j \in \mathcal{M}^S(x_j)$ and $y_j \rightarrow y^*$, with $x_j, x^* \in \mathbb{O}^{n \times k}, y_j, y^* \in \mathbb{R}^{p \times k}$. Suppose $y^* \notin \mathcal{M}^S(x^*)$. Since $y^* \in \mathbb{R}^{p \times k}$, $F(y^*, x^*) > \min_{\hat{y} \in \mathbb{R}^{p \times k}} F(\hat{y}, x^*) = L$. There exists some $\epsilon_0 > 0$ such that $F(y^*, x^*) > L + \epsilon_0$. Let $\tilde{y} \in \mathcal{M}^S(x^*)$. Then $F(\tilde{y}, x^*) = L$. Since $F$ is continuous at $(\tilde{y}, x^*)$, $\lim_{j \rightarrow \infty} F(\tilde{y}, x_j) = L$. For $j$ large enough, $|F(\tilde{y}, x_j) - L| \leq \epsilon_0 / 2$, from which it follows that $F(y_j, x_j) \leq F(\tilde{y}, x_j) \leq L + \epsilon_0 / 2$ and $F(y^*, x^*) \leq L + \epsilon_0 / 2$.
The contradiction implies \( y^* \in \mathcal{M}^s(x^*) \). Hence \( \mathcal{M}^s \) and thus \( \mathcal{M}^S \) are closed. From the proof of Lemma 12, we also know \( \mathcal{M}^S \) is compact. Similarly, we can show \( \mathcal{M}^V \) is closed on \( \mathbb{R}^{p \times k} \). Based on the properties of the point-to-set maps (Luenberger and Ye, 2008, p. 205), \( \mathcal{M} \) is closed on \( \Omega \).

If we set \( Z \) in the general statement of the Global Convergence Theorem to be our continuous criterion function \( F \), and if we take \( \Gamma \) to be the set of local of minima of \( F \), Lemma 11 and the assumption in Part (ii) of our Theorem 7 guarantee that (2) of GCT holds. Lemmas 12 and 13 verify conditions (1) and (3) of GCT, respectively. This concludes the proof of this part.

**Proof of Part (i)** From displays (25) and (26) we observe that \( F \) is **convex** in \( S \), given \( V \), and it is linear and therefore **smooth** in \( V \), given \( S \). This part of the theorem shows that, under no further conditions on \( F \), algorithm \( \mathcal{A} \) converges to a stationary point of \( F \). We begin by defining a stationary point in this context. Recall that we view (13) as an unconstrained optimization problem in \( \Omega = \mathbb{R}^{p \times k} \times \mathbb{O}^{n \times k} \). Notice that \( \mathbb{O}^{n \times k} \), which is a Stiefel manifold, is a Riemannian submanifold of \( \mathbb{R}^{n \times k} \). We use the inherited Riemannian metric to define the gradient of \( F \) with respect to \( V \) (Boothby, 2002). This Riemannian gradient, denoted by \( \nabla_V F(S, V) \), can be explicitly computed: 

\[
\nabla_V F(S, V) = \mathcal{P}_V(-W) = -W + VVV'/2 + VWV'/2 \text{ with } W = Y'XS,
\]

where \( \mathcal{P}_V \) is the projection onto the tangent space to \( \mathbb{O}^{n \times k} \) at \( V \).

Since \( F \) is convex in \( S \), the subdifferential of \( F \) with respect to \( S \), denoted by \( \partial_S F \), is also well defined. From Gabay (1982) and Shimizu, Ishizuka, and Bard (1997, p. 62), a necessary condition for \( F \) to have a local minimum at \((S^*, V^*) \) is that this point is
stationary, that is

$$0 \in \partial_S F(S^*, V^*), \quad \text{and} \quad \nabla_V F(S^*, V^*) = 0. \quad (28)$$

For future use, note that $\nabla_V F(S, V)$ is continuous on $\Omega$. Since $V^{(j+1)}$ minimizes $F(S^{(j+1)}, \cdot)$, we have $\nabla_V F(S^{(j+1)}, V^{(j+1)}) = 0$, for any $j \geq 0$.

Because the optimum of (25) (or (26)) may not be uniquely attained, $F$ is not guaranteed to be a descent function in general (see Zangwill (1969), Bertsekas (1999) and Luenberger and Ye (2008) for details). Therefore, GCD cannot be directly applied.

From Lemma 11, $F(S^{(j)}, V^{(j)})$ must converge. Denote the limit by $L^*$. Let $(S^{(ji)}, V^{(ji)})$ ($l = 1, 2, \ldots$) be a subsequence of $(S^{(j)}, V^{(j)})$ which converges to $(S^*, V^*)$ as $l \to \infty$. Then $L^* = \lim_{l \to \infty} F(S^{(ji)}, V^{(ji)}) = F(S^*, V^*)$. We assume, without loss of generality, $(S^{(jl+1)}, V^{(jl+1)})$ also converges (Lemma 12), and denote the limit by $(\tilde{S}, \tilde{V})$. We have

$$F(\tilde{S}, \tilde{V}) = L^*. \quad (29)$$

We claim that $(S^*, V^*)$ must be a stationary point of $F$. First, the continuity of $\nabla_V F(S, V)$ and the fact that $\nabla_V F(S^{(j)}, V^{(j)}) = 0$ imply that $\nabla_V F(S^*, V^*) = 0$. Suppose $0 \notin \partial_S F(S^*, V^*)$. This implies $S^*$ is not a global minimizer of (25). Lemma 13, however, states that $(\tilde{S}, \tilde{V}) \in \mathcal{M}(S^*, V^*)$. That is, $L^* = F(\tilde{S}, \tilde{V}) \leq F(\tilde{S}, V^*) < F(S^*, V^*) < L^*$, which contradicts (29). The last inequality is strict because $F$ is convex in $S$ and so, applying the algorithm to $S^*$, which is not a global minimizer, yields a strict improvement (decrease) of the criterion function. Hence $0 \in \partial_S F(S^*, V^*)$. The proof is complete.  \qed
4.6 Proof of Theorem 8

We use the same notation system as in Appendix 4.5. Recall that \((S^*, V^*) \in \Omega\) is a coordinatewise minimum point of \(F(S, V)\) if \(F(S, V^*) \geq F(S^*, V^*), \forall S \in \mathbb{R}^{p \times k}\) and \(F(S^*, V) \geq F(S^*, V^*), \forall V \in O^{n \times k}\) (Tseng, 2001). In our problem, this implies \((S^*, V^*)\) is also a stationary point of \(F(S, V)\).

Without loss of generality, assume \(X\) has been scaled to have \(\|X\|_2 \leq 1\) before running Algorithm \(\mathcal{A}'\). For simplicity, set \(K = 1\) in (14) and redefine the operator \(T_V : \mathbb{R}^{p \times k} \rightarrow \mathbb{R}^{p \times k}\) by

\[
T_V \circ S = \Theta(X'YV + (I - X'X)S; \lambda), \forall S \in \mathbb{R}^{p \times k}.
\]  

(30)

Let \(T^\alpha_V (\alpha \in \mathbb{N})\) be the composition of \(\alpha\) \(T_V\)'s. Define point-to-set maps \(\mathcal{M}^S(S, V) = \{(\tilde{S}, \tilde{V}) \in \Omega : \tilde{S} \in \{T_V \circ S, T^2_V \circ S, \ldots, T^{M_{iter}}_V \circ S\}, \tilde{V} = V\}\), and \(\mathcal{M}^V(S, V) = \{(\tilde{S}, \tilde{V}) \in \Omega : \inf_{\tilde{V} \in O^{n \times k}} F(S, \tilde{V}) \geq F(S, \tilde{V}), \tilde{S} = S\}\). Then \(\mathcal{M} = \mathcal{M}^V \mathcal{M}^S\) characterizes \(\mathcal{A}'\).

When updating \(S\) at Step (a), the algorithm allows one to perform \(T\) any times (denoted by \(\alpha_j\)) provided \(\alpha_j\) does not go beyond \(M_{iter} \in \mathbb{N}\) that is prespecified before running \(\mathcal{A}'\).

**Lemma 14.** Given any \(V \in O^{n \times k}\) and \(S \in \mathbb{R}^{p \times k}\), let \(\tilde{S} = T_V(S)\). Then \(F(S, V) - F(\tilde{S}, V) \geq \frac{1}{2}\|\tilde{S} - S\|_F^2\).

**Proof.** Apply the theorem in She (2010) to the vectorized problem (25). Note that \(2 - \|X \otimes I\|_2 = 2 - \|X\|_2 \geq 1\). The proof details are omitted. \(\square\)

Choose \((\tilde{S}, V) \in \mathcal{M}^S(S, V)\), using the triangle inequality we know \(F(S, V) - F(\tilde{S}, V) \geq \frac{1}{2}\|\tilde{S} - S\|_F^2\).
Lemma 15. Suppose $\lambda > 0$. Then $B^{(j)}$, $S^{(j)}$, $V^{(j)}$ in $A'$ are uniformly bounded in $j$.

Proof. The proof is similar to the proof of Lemma 12 and is thus omitted.

Lemma 16. Suppose $\lambda > 0$. The $\mathcal{M}$ for describing $A'$ is a closed point-to-set map on $\Omega$.

Proof. Similar to the proof of Lemma 13, we prove that the point-to-set map $\mathcal{M}^s(S, V) = \{T_V \circ S, T_V^2 \circ S, \cdots, T_V^{\text{iter}} \circ S\}$ is closed at any $(S, V) \in \Omega$. Then $M^S$ and thus $\mathcal{M}$ are closed on $\Omega$.

Let $(S_j, V_j) \rightarrow (S^*, V^*)$, $\tilde{S}_j \in \mathcal{M}^s(S_j, V_j)$ and $\tilde{S}_j \rightarrow \tilde{S}^*$, with $(S_j, V_j), (S^*_j, V^*_j) \in \Omega$ and $\tilde{S}_j, \tilde{S}^* \in \mathbb{R}^{p \times k}$. There must exist infinitely many $\tilde{S}_j$'s satisfying $T_{V_j}^\alpha \circ (S_j, V_j) = \tilde{S}_j$ for some $\alpha \in \mathbb{N}$. Let $g(S, V) = T_{V}^\alpha(S, V)$. It is not difficult to see that $g$ is jointly continuous. Hence $\tilde{S}^* \in \mathcal{M}^s(S^*, V^*)$ by a subsequence argument.

Now we prove Theorem 8. Following the lines of the proof of Part (i) of Theorem 7, for any accumulation point $(S^*, V^*)$ of $(S^{(j)}, V^{(j)})$, $(S^*, V^*) \in \Omega$ and there exists $(\tilde{S}, \tilde{V}) \in \mathcal{M}(S^*, V^*)$ with $F(\tilde{S}, \tilde{V}) = F(S^*, V^*)$. Since $F(\tilde{S}, \tilde{V}) \leq F(S^*, V^*) \leq F(S^*, V^*)$, $F(\tilde{S}, V^*) = F(S^*, V^*)$. It follows from the comment after Lemma 14 that $\tilde{S} = S^*$. This means $T_{V^*}^{\alpha_0} \circ S^* = S^*$ for some $\alpha_0 \in \mathbb{N}$. But then $F(T_{V^*}^{\alpha} \circ S, V^*) = F(S^*, V^*)$ for any $\alpha \leq \alpha_0$, and in particular, $F(T_{V^*} \circ S, V^*) = F(S^*, V^*)$. Applying Lemma 14 again yields $T_{V^*} \circ S^* = S^*$. It is easy to verify from (30) that $S^*$ is a fixed point of $T_{V^*}$ is equivalent to $0 \in \partial_S F(S^*, V^*)$. Therefore, $S^*$ is a global minimizer of $F(S, V^*)$ given $V^*$, due to the convexity of (25).

On the other hand, from $\tilde{S} = S^*$, we have $(S^*, \tilde{V}) \in \mathcal{M}^V(S^*, V^*)$. $\tilde{V}$ is a (global) minimizer of $F(S^*, V)$ given $S^*$. But $F(S^*, \tilde{V}) = F(\tilde{S}, \tilde{V}) = F(S^*, V^*)$, so $V^*$ also
minimizes $F(S^*, V)$ given $S^*$, and $\nabla_V F(S^*, V^*) = 0$. In summary, $(S^*, V^*)$ is a coordinatewise minimum of $F$. □

References


