JOINT VARIABLE AND RANK SELECTION FOR PARSIMONIOUS ESTIMATION OF HIGH-DIMENSIONAL MATRICES

By Florentina Bunea[1], Yiyuan She[2] and Marten H. Wegkamp[1]

Cornell University, Florida State University and Cornell University

We propose 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. We 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. Next, 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 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

$$Y = XA + E \tag{1}$$

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 $m \times n$ matrix with independent $N(0, \sigma^2)$ entries. 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 denote the rank of A and J denote the index set of the nonzero rows of A and $|J|$ 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, 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 consists of *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. 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.

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 embodies both sparsity constraints on A simultaneously. In this work we introduce *row and rank sparse* models, and suggest and analyze new methods that combine the strengths of the existing dimension reduction techniques. We propose penalized least squares methods, with new penalties tailored to adaptive and optimal estimation in the *row and rank sparse* model (1). The rest of the article is organized as follows.

We introduce in Section 2.1 a product-type penalty that imposes simultaneously rank and row sparsity restrictions on the coefficient matrix. It generalizes both AIC-type penalties developed for variable selection in univariate response regression models as well as the rank penalty of Bunea, She and Wegkamp (2011) for low rank estimation in multivariate response models. The purpose of the resulting method is twofold. 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. Their rates of convergence coincide with the existing minimax rates in the literature, up to a logarithmic term; cf. Koltchinskii, Lounici and Tsybakov (2011). Second, we show in Theorem 2 that this method can also be employed for selecting among competing estimators from a large finite 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 m, but computing the estimator analyzed in Theorem 1 requires an exhaustive search over the class of all possible models, the size of which is exponential in p, and this becomes computationally prohibitive if $p > 20$.

To address the computational issue, 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 estimators of A. Both are computable in high dimensions, and both achieve the rates discussed in Section 2.1, up to a $\log(p)$ factor, 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 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 principal components analysis (PCA) and canonical correlation analysis (CCA) are perhaps the most popular ways of achieving dimension reduction of multivariate data. 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 known as *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 penalized least squares estimators. For penalties proportional to the nuclear norm, we refer to Yuan et al. (2007), Candès and Plan (2010), Negahban and Wainwright (2011), Rohde and Tsybakov (2011). For penalties proportional to the rank, we refer to Bunea, She and Wegkamp (2011) and Giraud (2011). Both types of estimators are computationally efficient, even if

TABLE 1
Oracle rate comparison between JRRS,
GLASSO and RSC. The sample size and
dimension parameters $m, p, n, q, r, |J|$ satisfy
$q \leq m \wedge p, \ r \leq n \wedge |J|, \ |J| \leq q$

GLASSO:	$n|J| + |J|\log(p)$
RSC or NNP:	$nr + qr$
JRRS:	$nr + |J|r\log(p/|J|)$

$\max(n, p) > m$, and both 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, for example, Rohde and Tsybakov (2011) for lower bound calculations.

To explain the other notion of sparsity, note that removing predictor X_j from model (1) is equivalent with setting the jth row in A to zero. Since vectorizing both sides of model (1) yields a univariate response regression model, we can view the rows of A as groups of coefficients in the transformed model. We can set them to zero by any group selection method developed for univariate response regression models in high dimensions such as the Group Lasso [Yuan and Lin (2006)], GLASSO for later reference. 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 et al. (2011) 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. For ease of reference, Table 1 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.

The table reveals that if $n \geq q$, the rates of the RSC, NNP and JRRS are dominated by nr, regardless of J, while if $n < q$, the new class of methods can provide substantial rate improvements over the existing methods, especially when the rank r is low.

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 (2011) to accommodate variable selection. We propose our single-stage joint rank and row selection (JRRS) estimator

$$(2) \qquad \hat{B} = \underset{B}{\arg\min}\{\|Y - XB\|_F^2 + \text{pen}(B)\},$$

also denoted by JRRS1, with penalty term

$$(3) \quad \text{pen}(B) = c\sigma^2 r(B)\left\{2n + \log(2e)|J(B)| + |J(B)|\log\left(\frac{ep}{|J(B)|}\right)\right\}.$$

The penalty is essentially proportional to the number of parameters in a model with fewer predictors $J(B)$ and of reduced rank $r(B)$. Here $c > 3$ is a numerical constant, $J(B)$ is the set of indices of nonzero rows, $r(B)$ is the rank of a generic $p \times n$ matrix B and the squared Frobenius norm of a generic matrix M is denoted by $\|M\|_F^2$ and is equal to the sum of the squared entries of M.

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 $|J|$ nonzero 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 (2) above achieves the lower bound, up to a log factor, under no restrictions on the design X, rank r or dimensions m, n, p.

THEOREM 1. *The single-stage JRRS estimator* $\hat{B}$ *in (2) using pen(B) in (3) with* $c = 12^3$ *satisfies*

$$\mathbb{E}[\|XA - X\hat{B}\|_F^2] \leq 10\|XA - XB\|_F^2 + 8\,\text{pen}(B) + 768n\sigma^2\exp(-n/2)$$

for any B with $r(B) \geq 1$. *In particular, if* $r(A) \geq 1$,

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

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

The proof of Theorem 1 remains valid if the matrices we select from depend on the data. Thus, our procedure can be used for selecting from any countable list of random matrices of different ranks and with different sparsity patterns. We will make essential use of this fact in the next section.

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

$$(4) \qquad\qquad \tilde{B} = \arg\min_{B_j}\|Y - XB_j\|_F^2 + \text{pen}(B_j)$$

---

[3]Our proof shows that we may take $c > 3$, at the cost of increasing numerical constants in the right-hand side of the oracle inequality.

with $c = 12$ satisfies

$$\mathbb{E}[\|XA - X\tilde{B}\|_F^2] \leq \inf_j \{10\mathbb{E}[\|XA - XB_j\|_F^2] + 8\mathbb{E}[\text{pen}(B_j)]\}$$

$$+ 768n\sigma^2 \exp(-n/2).$$

2.2. *Two-step joint rank and row selection estimators.* The computational complexity of the single-stage JRRS estimator (2) 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 (3) 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 alternatives, each a two-step JRRS procedure and 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

(5) $$\hat{B}_k = \underset{r(B) \leq k}{\arg\min} \{\|Y - XB\|_F^2 + 2\lambda\|B\|_{2,1}\}.$$

Here λ is a tuning parameter and the minimization is over all $p \times n$ matrices B of rank less than or equal to (a fixed) k. A computationally efficient numerical algorithm for solving this minimization problem is given in Section 3.1. 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 $\hat{B}_k$ as the rank constrained group lasso (RCGL) estimators.

Since this estimator is central to our procedures, we analyze it first. We need the following mild assumption on $\Sigma = X'X/m$.

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

(6) $$\text{tr}(M'\Sigma M) \geq \delta_I \sum_{i \in I} \|m_i\|_2^2$$

for all $p \times n$ matrices M (with rows m_i) satisfying $\sum_{i \in I} \|m_i\|_2 \geq 2\sum_{i \in I^c} \|m_i\|_2$.

REMARKS. (1) The constant 2 may be replaced by any constant larger than 1.

(2) Assumption $\mathfrak{A}$ allows designs with $p > m$, and can be seen as a version of the restricted eigenvalue condition in the variable selection literature introduced in Bickel, Ritov and Tsybakov (2009) and analyzed in depth in Bühlmann and van de Geer (2011).

(3) A sufficient condition for (6) is: there exists a diagonal matrix D with $D_{jj} = \delta_I$ for all $j \in I$ and $D_{jj} = 0$ otherwise such that $\Sigma - D$ is positive definite.

Let $\lambda_1(\Sigma)$ denote the largest eigen-value of Σ and set the tuning parameter

$$(7) \qquad \lambda = C\sigma\sqrt{\lambda_1(\Sigma)km\log(ep)}$$

for some numerical constant $C > 0$. Notice that λ depends on k, but we suppress this dependence in our notation.

THEOREM 3. *Let $\hat{B}_k$ be the global minimizer of (5) corresponding to λ in (7) with C large enough. Then, we have*

$$(8) \quad \begin{aligned} \mathbb{E}[\|X\hat{B}_k - XA\|_F^2] &\lesssim \|XB - XA\|_F^2 \\ &\quad + k\sigma^2\left\{n + \left(1 + \frac{\lambda_1(\Sigma)}{\delta_{J(B)}}\right)|J(B)|\log(p)\right\} \end{aligned}$$

for any $p \times n$ matrix B with $1 \le r(B) \le k$, $|J(B)|$ nonzero rows, provided Σ satisfies assumption $\mathfrak{A}(J(B), \delta_{J(B)})$.

REMARKS. (1) The term $\{n + |J(B)|\log(p)\}k\sigma^2$ in (8) is multiplied by a factor $\kappa = \lambda_1(\Sigma)/\delta_{J(B)}$. This factor can be viewed as a generalized condition number of the matrix Σ. If κ stays bounded, 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.

(2) 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 λ markedly differs from the choice proposed by Lounici et al. (2011) for the GLASSO estimator $\hat{A}$. We need a different choice for λ and a more refined analysis since we minimize in (5) over all $p \times n$ matrices B of rank $r(B) \le k$.

2.2.2. *Adaptive rank-constrained predictor selection.* We now develop theoretical properties of three methods, Method 1 (RSC→RCGL), Method 2 (RCGL→JRRS1) and Method 3 (GLASSO→RSC). *JRRS1* denotes the single-stage JRRS estimator of Section 2.1.

Theorem 3 suggests that by complementing RCGL by a method that estimates the rank consistently, we could obtain row and rank optimal adaptive estimator. This is indeed true.

METHOD 1 (RSC→RCGL).

- Use the rank selection criterion (RSC) of Bunea, She and Wegkamp (2011) to select $k = \hat{r}$ as the number of singular values of PY that exceed $\sigma(\sqrt{2n} + \sqrt{2q})$. Here P is the projection matrix on the space spanned by X.

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

This two-step estimator adapts to both rank and row sparsity, under two additional, mild restrictions.

ASSUMPTION $\mathfrak{C}$1. $d_r(XA) > 2\sqrt{2}\sigma(\sqrt{n} + \sqrt{q})$.
ASSUMPTION $\mathfrak{C}$2. $\log(\|XA\|_F) \le (\sqrt{2} - 1)^2(n + q)/4$.

Assumption $\mathfrak{C}$1 only requires that the signal strength, measured by $d_r(XA)$, the rth singular value of the matrix XA, be larger than the "noise level" $2\sigma\sqrt{2n + 2q}$, otherwise its detection would become problematic. The tightness of $\mathfrak{C}$1 is discussed in detail in Bunea, She and Wegkamp (2011). Theorem 2 of that work proves that the correct rank will be selected with probability[4] $1 - \exp\{-c_0(n + q)\}$ with $c_0 = (\sqrt{2} - 1)^2/2$.

Assumption $\mathfrak{C}$2 is technical and needed to guarantee that the error due to selecting the rank is negligible compared to the rate $nr + |J|r\log(p)$.

THEOREM 4.   *Let Σ satisfy $\mathfrak{A}(J, \delta_J)$ with $J = J(A) \ne \varnothing$, let $\lambda_1(\Sigma)/\delta_J$ be bounded, and let $\mathfrak{C}$1 and $\mathfrak{C}$2 hold. Then the two-step JRRS estimator $\hat{B}^{(1)}$ with λ set according to (7) with C large enough satisfies*

$$\mathbb{E}[\|X\hat{B}^{(1)} - XA\|_F^2] \lesssim nr + |J|r\log(p).$$

Hence, $\hat{B}^{(1)}$ is *row and rank* adaptive, and achieves the same optimal rate, up to a $\log(p)$ factor, as the *row and rank* adaptive $\hat{B}$ studied in Theorem 1 above. While Theorem 1 is proved under no restrictions on the design, we view the mild conditions of Theorem 4 as a small price to pay for the computational efficiency of $\hat{B}^{(1)}$ relative to that of $\hat{B}$ in (2). The practical choice of the threshold $2\sigma\sqrt{2n + 2q}$ in the initial step of our procedure can be done either by replacing σ^2 by an estimator, as suggested and analyzed theoretically in Section 2.4 of Bunea, She and Wegkamp (2011), 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 (2011).

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 $\mathfrak{C}$1 and $\mathfrak{C}$2.

METHOD 2 (RCGL→JRRS1).

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

---

[4]Hence, if $n + q$ is small compared to m, we suggest to replace q by $q\log(m)$ in the threshold level in the definition of $\hat{r}$, in $\mathfrak{C}$1 and in $\mathfrak{C}$2; see the remark following Corollary 4 in Bunea, She and Wegkamp (2011).

- Compute $\hat{B}^{(2)} = \arg\min_{B \in \mathcal{B}} \{ \|Y - XB\|_F^2 + \text{pen}(B) \}$, with $\text{pen}(B)$ defined in (3) above.

We have the same conclusion as for Method 1:

THEOREM 5. *Provided Σ satisfies condition $\mathfrak{A}(J, \delta_J)$ with $J = J(A) \neq \varnothing$, $\lambda_1(\Sigma)/\delta_J$ is bounded, and Λ contains λ in (7) for some C large enough, we have*

$$\mathbb{E}[\|X\hat{B}^{(2)} - XA\|_F^2] \lesssim nr + |J| \log(p) r.$$

We see that $\hat{B}^{(2)}$ has the same rate as $\hat{B}^{(1)}$, under condition $\mathfrak{A}$ on the design only. Our simulation studies in Section 4 indicate that the numerical results of Methods 1 and 2 are comparable.

REMARK. A perhaps more canonical two-stage procedure is as follows:

METHOD 3 (GLASSO→RSC).

- Select the predictors via the GLASSO.
- Use the rank selection criterion (RSC) of Bunea, She and Wegkamp (2011) to construct an adaptive estimator, of reduced rank, based only on the selected predictors.

It is clear that as soon as we have selected 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 (2011). 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 mutual coherence condition on $\Sigma = (\Sigma_{ij})_{1 \leq i,j \leq p}$, which postulates that the off-diagonal elements of Σ be small. Specifically, for the GLASSO, the restriction becomes $|\Sigma_{ij}| \leq 1/(7\alpha|J|)$, for some $\alpha > 1$ [cf. Lounici et al. (2011)], if it is coupled with the condition that $n^{-1/2}\|a_j\| \geq C_1 m^{-1/2}[1 + C_2 n^{-1/2} \log p]^{1/2}$. Here $\|a_j\|$ is the Euclidean norm of the jth row vector of A, and C_1 and C_2 are constants. For designs for which Σ 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 $n^{-1/2}\|a_j\| \geq C_1 m^{-1/2}[1 + C_2 n^{-1} \log p]^{1/2}$; see Corollary 5.2 in Lounici et al. (2011). Our Theorems 4 and 5 require substantially weaker assumptions on the design.

Algorithm $\mathscr{A}$ Rank constrained group lasso (RCGL) computation

---

given $1 \leq k \leq m \wedge p \wedge n$, $\lambda \geq 0$, $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n \times k}$ (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: $W = U_w D_w V_w'$.
   (c) $V_{k,\lambda}^{(j+1)} \leftarrow U_w V_w'$
   (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 $\hat{B}_{k,\lambda} = B_{k,\lambda}^{(j+1)}$, $\hat{S}_{k,\lambda} = S_{k,\lambda}^{(j+1)}$, $\hat{V}_{k,\lambda} = V_{k,\lambda}^{(j+1)}$.

---

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

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

over all $p \times n$ matrices B of rank less than or equal to k. Recall that by solving this problem we provide a way of performing *rank-constrained variable selection* in model (1). Directly solving the nonconvex constrained minimization problem for B in (9) 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

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

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 (10) by $F(S, V; \lambda)$. We propose the following iterative optimization procedure.

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

THEOREM 6.   *Given $\lambda > 0$ and an arbitrary starting point $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n \times k}$, let $(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ $(j = 1, 2, \ldots)$ be the sequence of iterates generated by Algorithm $\mathscr{A}$. 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 $(S_{k,\lambda}^*, V_{k,\lambda}^*)$.*

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

REMARKS. (1) We run the algorithm to obtain a solution path, for each (k, λ) in a two-dimensional grid or for a grid of λ with k determined by RSC. From the solution path, we get a series of candidate estimates. Then the single stage JRRS (4) or other tuning criteria can be used to select the optimal estimate.

(2) Our results are of the same type as those established for the convergence of the EM algorithm [Wu (1983)]. Algorithm $\mathscr{A}$ can be viewed as a block coordinate descent method, but the conclusion in Theorem 6 is stronger in some sense: the guaranteed convergence to a stationary point (to be defined in Appendix A.6) does *not* require the uniqueness of $S_{k,\lambda}^{(j+1)}$ in step (a) which is a crucial assumption in the literature [see, e.g., Bertsekas (1999) and Tseng (2001)].

(3) Step (a) needs to solve a GLASSO optimization problem. To see this, denoting the standard vectorization operator by vec and the Kronecker product by $\otimes$, we rewrite $\|YV_k - XS_k\|_F^2/2 + \lambda\|S_k\|_{2,1}$ as $\|\text{vec}((YV_k)') - (X \otimes I_k)\text{vec}(S_k')\|_F^2/2 + \lambda\|S_k\|_{2,1}$. Although this subproblem is convex, finding its global minimum point can still be expensive for large data. Instead, one may perform some low-cost thresholding for a few steps. Concretely, let K be a constant satisfying $K > \|X\|_2^2/2$. Given $V \in \mathbb{O}^{p \times k}$, define $T_V : \mathbb{R}^{p \times k} \to \mathbb{R}^{p \times k}$ as

$$(11) \quad T_V \circ S = \vec{\Theta}\left(\frac{1}{K}X'YV + \left(I - \frac{1}{K}X'X\right)S; \frac{\lambda}{K}\right) \qquad \forall S \in \mathbb{R}^{p \times k},$$

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

We now replace step (a) in Algorithm $\mathscr{A}$ 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 α_j, satisfies $1 \leq \alpha_j \leq M_{\text{iter}}$ for some $M_{\text{iter}} < \infty$ specified based on available computational resources. α_j need not be equal. This algorithm, denoted by $\mathscr{A}'$, offers more flexibility and is more convenient than Algorithm $\mathscr{A}$ in implementation. Although at each

iteration $S_{k,\lambda}^{(j+1)}$ is not uniquely determined, a stronger global convergence result holds for $\mathscr{A}'$.

THEOREM 7. *Given $\lambda > 0$ and an arbitrary starting point $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n \times k}$, let $(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ $(j = 1, 2, \ldots)$ be the sequence of iterates generated by $\mathscr{A}'$. 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 $F(S_{k,\lambda}^{(j)}, V_{k,\lambda}^{(j)})$ 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 MVN$(\mathbf{0}, \Sigma)$, with $\Sigma_{jk} = \rho^{|j-k|}$, $\rho > 0$, $1 \le j, k \le p$.
- The coefficient matrix A has the form

$$A = \begin{bmatrix} A_1 \\ O \end{bmatrix} = \begin{bmatrix} bB_0 B_1 \\ O \end{bmatrix}$$

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

This setup contains many noisy features, but the relevant features lie in a low-dimensional subspace. This structure resembles many real world data sets; 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, say, 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 suboptimal 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 data set

TABLE 2
Performance comparisons between GLASSO, RSC,
Methods 1, 2 and 3 in the $p > m$ experiment with
$b = 0.5, 1, |J| = 15$ *and* $r = 2$

	MSE	$|\hat{J}|$	$\hat{R}$	M	FA
$b = 0.5$					
GLASSO	206	10	10	53%	4%
RSC	485	100	2	0%	100%
Method 1	138	19	2	36%	10%
Method 2	169	21	2	45%	7%
Method 3	169	10	2	53%	4%
$b = 1$					
GLASSO	511	14	10	41%	7%
RSC	1905	100	2	0%	100%
Method 1	363	21	2	31%	12%
Method 2	434	25	2	30%	13%
Method 3	402	14	2	41%	7%

(10,000 observations) to tune the parameter of each algorithm (with the exception of Method 2) and we also generated another independent data set of the same size as the test data to evaluate the test error. Similar to the LARS-OLS hybrid [Efron et al. (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 $c = 3$ and $\sigma^2 = 1$ in (3) 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) $\|XA - X\hat{A}\|_F^2/(mn)$ 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, and it therefore allows for a more fair comparison between methods.

We also report the median number of predictors (denoted by $|\hat{J}|$) and median rank estimate (denoted by $\hat{R}$) over all runs. Estimators with small MSE and low $|\hat{J}|$ and $\hat{R}$ are preferred from the point of view of statistical modeling.

Finally, we provide the rates of nonincluded true variables (denoted by M for misses) and the rates of incorrectly included variables (FA for false

TABLE 3
Performance comparisons between GLASSO, RSC,
Methods 1, 2 and 3 in the $m > p$ experiment with
$b = 0.2, 0.4$, $|J| = 15$ and $r = 5$

	MSE	$|\hat{J}|$	$\hat{R}$	M	FA
		$b = 0.2$			
GLASSO	18.1	14	14	4%	1%
RSC	11.9	25	5	0%	100%
Method 1	8.3	15	5	0%	1%
Method 2	8.3	15	5	0%	4%
Method 3	8.9	14	5	4%	1%
		$b = 0.4$			
GLASSO	17.7	15	15	0%	0%
RSC	11.5	25	5	0%	100%
Method 1	8.1	15	5	0%	0%
Method 2	8.0	15	5	0%	1%
Method 3	8.1	15	5	0%	0%

alarms). Ideally, both rates are low, especially the M-rates, since we do not wish to discard relevant features.

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 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 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 Methods 1 and 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 (3) 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 Methods 1 and 2 in the $m > p$ experiment, when the 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 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 (3) provides evidence of success of Theorem 2.

4. Applications. In this section we apply Method 1, with its tuning parameters chosen via cross-validation, to two real data sets 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 nonlinear 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 (2011), 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

TABLE 4
Paper quality control: joint new feature construction from the selected predictors with
$\hat{r} = 3, |\hat{J}| = 6$. *The extracted components are ordered by their associated eigenvalues*

New scores	Eigenvalues	X_1	X_2	X_3	X_2^2	X_3^2	$X_1 \cdot X_3$
1	112.4	1.9244	-0.5288	-1.2321	-0.4443	1.3109	1.1898
2	40.7	0.8231	-0.5937	0.9324	0.7819	-0.1599	-0.8536
3	24.9	0.2871	1.0336	0.4215	0.8365	0.4245	0.4677

Reinsel and Velu (1998) and Izenman (2008) for the explanation]. For instance, the first important 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 $2X_1 + X_1X_3 + 1_{X_3=-1} - 1_{X_2=1}$. 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.

Cognitive neuroimaging. We present an analysis of the data set described in Bunea et al. (2011) 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 diffusion 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 $\times$ DTI interactions. We refer to Bunea et al. (2011) 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.

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.

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 not only demonstrates the existence of a strong association between the variable `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 variable fractional anisotropy at corpus callosum (`fa_cc1`) stands out among the very many DTI-derived measures, in terms of predictive power.

APPENDIX

For a generic index set $I \subseteq \{1, \ldots, p\}$, we define the $m \times p$ matrix X_I as follows: its ith column coincides with that of X if $i \in I$, otherwise we set the entire column to zero. Furthermore, we define P_I as the projection matrix on the column space of X_I.

Since we favor transparent proofs, we did not attempt to optimize various numerical constants.

A.1. Proof of Theorem 1. By the definition of $\hat{B}$, for any $p \times n$ matrix B with $r(B) \geq 1$, the inequality

$$\|Y - X\hat{B}\|_F^2 + \text{pen}(\hat{B}) \leq \|Y - XB\|_F^2 + \text{pen}(B)$$

holds. This is equivalent with

$$
\begin{aligned}
\text{(12)} \quad \|XA - X\hat{B}\|_F^2 \leq \|XA - XB\|_F^2 &+ 2\,\text{pen}(B) \\
&+ \{2\langle E, X\hat{B} - XB\rangle - \text{pen}(\hat{B}) - \text{pen}(B)\}.
\end{aligned}
$$

We consider two complementary cases: $r(\hat{B}) \geq 1$ and $r(\hat{B}) = 0$.

Case 1: $r(\hat{B}) \geq 1$. We write $J = J(B)$ and $\hat{J} = J(\hat{B})$, and we note that $J \cup \hat{J} = J_1 \cup J_2 \cup J_3$ with $J_1 = \hat{J}^c \cap J$, $J_2 = J \cap \hat{J}$, $J_3 = J^c \cap \hat{J}$. Hence,

$$\langle E, X(\hat{B} - B)\rangle = \sum_{k=1}^{3} \langle E, X_{J_k}(\hat{B} - B)\rangle.$$

The penalty term in (3) can be written as

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

for the function $f(x) = x\log(2e) + x\log(ep/x)$. Since $f''(x) = -1/x$, $f(x)$ is concave for $x > 0$, and we have

$$f(x + y) \geq \frac{1}{2}f(2x) + \frac{1}{2}f(2y) = x + y + x\log\left(\frac{ep}{x}\right) + y\log\left(\frac{ep}{y}\right)$$

for all $x, y > 0$. Consequently, writing $r_1 = r(B)$, $r_3 = r(\hat{B})$ and $r_2 = r_1 + r_3$,

$$\text{pen}(B) + \text{pen}(\hat{B}) = c\sigma^2 r_1\{2n + f(|J_1| + |J_2|)\} + c\sigma^2 r_3\{2n + f(|J_2| + |J_3|)\}$$

$$\geq c\sigma^2 \sum_{k=1}^{3} r_k \left\{ n + |J_k| + |J_k| \log\left(\frac{ep}{|J_k|}\right) \right\}.$$

This implies that

$$\{2\langle E, X\hat{B} - XB\rangle - \text{pen}(\hat{B}) - \text{pen}(B)\}$$

$$\leq \sum_{k=1}^{3} \left[2\langle E, X_{J_k}(\hat{B} - B)\rangle - c\sigma^2 r_k \left\{ n + |J_k| + |J_k| \log\left(\frac{ep}{|J_k|}\right) \right\} \right].$$

We define

$$R = \max_{I} \left(d_1^2(P_I E) - \frac{c}{2}\sigma^2 \left\{ n + |I| + |I| \log\left(\frac{ep}{|I|}\right) \right\} \right)_{+}.$$

In this proof,[5] we set $c = 12$. Using the inequality

$$\langle E, X_{J_k}(\hat{B} - B)\rangle = \langle P_{J_k} E, X_{J_k}(\hat{B} - B)\rangle$$

$$\leq d_1(P_{J_k} E)\sqrt{r_k}\|X_{J_k}(\hat{B} - B)\|_F$$

and the inequalities $2xy \leq x^2/a + y^2 a$ and $(x + y)^2 \leq x^2(1 + a) + y^2(1 + a)/a$ for all $x, y \in \mathbb{R}$ with $a = 2$, we further bound

$$\{2\langle E, X\hat{B} - XB\rangle - \text{pen}(\hat{B}) - \text{pen}(B)\}$$

$$\leq \sum_{k=1}^{3} \left[\frac{1}{2}\|X_{J_k}(\hat{B} - B)\|_F^2 + 2r_k d_1^2(P_{J_k} E) \right.$$

$$\left. - c\sigma^2 r_k \left\{ n + |J_k| + |J_k| \log\left(\frac{ep}{|J_k|}\right) \right\} \right]$$

$$\leq \frac{1}{2}\|X\hat{B} - XB\|_F^2 + 6\left(\max_k r_k\right) R$$

$$\leq \frac{3}{4}\|X\hat{B} - XA\|_F^2 + \frac{3}{2}\|XB - XA\|_F^2 + 12nR,$$

using $\max(r_1, r_2, r_3) = r_2 \leq 2n$. Now, (12) and the display above yield

$$\tfrac{1}{4}\mathbb{E}[\|X\hat{B} - XA\|_F^2] \leq \tfrac{5}{2}\|XB - XA\|_F^2 + 2\,\text{pen}(B) + 12n\mathbb{E}[R]$$

$$\leq \tfrac{5}{2}\|XB - XA\|_F^2 + 2\,\text{pen}(B) + 192n\sigma^2 \exp(-n/2)$$

using Lemma 8 in the last inequality. This concludes the first case.

---

[5]A careful inspection reveals that we may take any $c > 3$, at the cost of larger constants elsewhere.

Case 2: $r(\hat{B}) = 0$. Using the same reasoning as above, we can argue that

$$\|XA\|_F^2 = \|XA - X\hat{B}\|_F^2$$
$$\leq \|XB - XA\|_F^2 + 2\operatorname{pen}(B) - 2\langle E, XB\rangle - \operatorname{pen}(B)$$
$$\leq \tfrac{5}{2}\|XB - XA\|_F^2 + 2\operatorname{pen}(B) + \tfrac{3}{4}\|XA\|_F^2 + R'$$

with

$$R' = 2r(B)d_1^2(P_{J(B)}E) - \operatorname{pen}(B)$$
$$\leq 2r(B)\left\{d_1^2(P_{J(B)}E) - \frac{c}{2}\sigma^2(n + |J(B)|)\right\}.$$

By Lemma 3 in Bunea, She and Wegkamp (2011), we have $\mathbb{E}[d_1^2(P_{J(B)}E)] \leq 2\sigma^2(n + |J(B)|)$, so that $\mathbb{E}[R'] \leq 0$ for our choice $c = 12$ above. Hence,

$$\tfrac{1}{4}\mathbb{E}[\|XA - X\hat{B}\|_F^2] \leq \tfrac{5}{2}\|XB - XA\|_F^2 + 2\operatorname{pen}(B),$$

which concludes the second case, and our proof.

LEMMA 8. *We have*

$$\mathbb{E}\left[\max_J\left(d_1^2(P_J E) - 6\sigma^2\left\{n + |J| + |J|\log\left(\frac{ep}{|J|}\right)\right\}\right)_+\right] \leq 16\sigma^2\exp(-n/2).$$

PROOF. Notice that $d_1^2(P_J E)$ has the same distribution as $d_1^2(Z_k)$ for a $m \times k$ matrix Z_k of independent $N(0, \sigma^2)$ entries with $k = |J|$. Consequently, for any $\nu_k \geq 0$,

$$\mathbb{E}\left[\max_J(d_1^2(P_J E) - \sigma^2\nu_{|J|}^2)_+\right] \leq \sum_{k=1}^q \binom{p}{k}\mathbb{E}[(d_1^2(Z_k) - \sigma^2\nu_k^2)_+].$$

We write $V_k = d_1(Z_k)/\sigma$ and $\mu_k = \sqrt{n} + \sqrt{k}$, $1 \leq k \leq p$. From Lemma 3 in Bunea, She and Wegkamp (2011), we have

$$\mathbb{P}\{V_k - \mu_k \geq t\} \leq \exp(-t^2/2)$$

for all $t > 0$. Hence, for $\nu_k = \mu_k + \delta_k$ and $\delta_k > 0$, we obtain

$$\mathbb{E}[V_k^2 - \nu_k^2]_+ \leq \mathbb{E}[V_k^2 1\{V_k^2 \geq \nu_k^2\}]$$
$$\leq \int_{\mu_k + \delta_k}^\infty 2x\mathbb{P}\{V_k \geq x\}\,dx$$
$$\leq \int_{\delta_k}^\infty 2(\mu_k + s)\exp(-s^2/2)\,ds$$
$$= 2\left(1 + \frac{\mu_k}{\delta_k}\right)\exp(-\delta_k^2/2)$$

for $1 \leq k \leq p$. Choosing $\delta_k^2 = 2k \log(ep/k) + n + k$, we find

$$\mathbb{E}\Big[\max_J (d_1^2(P_J E) - \sigma^2 \nu_{|J|}^2)_+\Big] \leq \sigma^2 \sum_{k=1}^{q} \frac{(ep)^k}{k^k} 2\Big(1 + \frac{\mu_k}{\delta_k}\Big) \exp(-\delta_k^2/2)$$

$$\leq 6\sigma^2 \sum_{k=1}^{q} \frac{(ep)^k}{k^k} \exp(-\delta_k^2/2)$$

$$= 6\sigma^2 \sum_{k=1}^{q} \exp(-k/2) \exp(-n/2)$$

$$\leq 16\sigma^2 \exp(-n/2).$$

Since $\nu_k^2 \leq 2\delta_k^2 + 2\mu_k^2 \leq 4k \log(ep/k) + 6k + 6n$, the claim follows. $\square$

A.2. Proof of Theorem 2. By the same reasoning as in the proof of Theorem 1, we obtain

$$\frac{1}{4}[\|X\tilde{B} - XA\|_F^2] \leq \frac{5}{2}\|XB_j - XA\|_F^2 + 2\operatorname{pen}(B_j)$$

$$+ 12n \max_I \Big(d_1^2(P_I E) - \frac{c}{2}\sigma^2\Big(n + |I| + |I| \log \frac{ep}{|I|}\Big)\Big)_+$$

for any B_j, random or not, with $r(B_j) > 0$. Taking expectations on both sides of the inequality and applying Lemma 8 gives the result.

A.3. Proof of Theorem 3. We denote the row vectors of the matrix $\hat{B}_k$ by $\hat{b}_1, \ldots, \hat{b}_p$, and we write $\hat{J}_k = \{i : \hat{b}_i \neq 0\}$ for the index set of nonzero rows. Let B be any matrix with row vectors $b_1, \ldots, b_p \in \mathbb{R}^n$ with $r(B) \leq k$ and $J = \{j : b_j \neq 0\}$ such that $\Sigma = X'X/m$ satisfies condition $\mathfrak{A}(J, \delta_J)$. We use $\tilde{X}$ to denote $X_{\tilde{J}}$ with $\tilde{J} = \hat{J}_k \cup J$ and we write $\Delta^2 = \|XB - XA\|_F^2$ and $\hat{\Delta}_k^2 = \|X\hat{B}_k - XA\|_F^2$. In this notation, we have by the definition of $\hat{B}_k$

$$\hat{\Delta}_k^2 + 2\lambda\|\hat{B}_k\|_{2,1} \leq \Delta^2 + 2\langle E, X(\hat{B}_k - B)\rangle + 2\lambda\|B\|_{2,1}.$$

This implies that

(13)
$$\hat{\Delta}_k^2 + 2\lambda\|(B - \hat{B}_k)_{J^c}\|_{2,1}$$

$$\leq \Delta^2 + 2\langle E, \tilde{X}(\hat{B}_k - B)\rangle + 2\lambda\|(B - \hat{B}_k)_J\|_{2,1}.$$

For the second term on the right of (13), we note that

$$2\langle E, \tilde{X}(\hat{B}_k - B)\rangle = 2\langle P_{\tilde{J}} E, \tilde{X}(\hat{B}_k - B)\rangle$$

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

$$\leq 2d_1(P_{\tilde{j}}E)\sqrt{2k}(\hat{\Delta}_k + \Delta)$$

$$\leq 16kd_1^2(P_{\tilde{j}}E) + \tfrac{1}{4}(\hat{\Delta}^2 + \Delta^2)$$

using $r(\hat{B}_k - B) \leq r(\hat{B}_k) + r(B) \leq 2k$ and the inequality $2xy \leq 4x^2 + y^2/4$ for all $x, y \in \mathbb{R}$. This bound and inequality (13) give the inequality

(14)
$$\tfrac{3}{4}\hat{\Delta}_k^2 + 2\lambda\|(\hat{B}_k - B)_{J^c}\|_{2,1}$$
$$\leq \tfrac{5}{4}\Delta^2 + 16kd_1^2(P_{\tilde{j}}E) + 2\lambda\|(\hat{B}_k - B)_J\|_{2,1}.$$

For the remainder of the proof, we consider two complementary cases.

Case 1. *Assume that*

(15)
$$\tfrac{5}{4}\Delta^2 + 16kd_1^2(P_{\tilde{j}}E) \leq 2\lambda\|(\hat{B}_k - B)_J\|_{2,1}.$$

In this case, (14) implies that

$$\|(\hat{B}_k - B)_{J^c}\|_{2,1} \leq 2\|(\hat{B}_k - B)_J\|_{2,1}.$$

Since Σ satisfies $\mathfrak{A}(J, \delta_J)$, we have

$$\|X\hat{B}_k - XB\|_F^2 \geq m\delta_J \sum_{j \in J} \|\hat{b}_i - b_i\|_2^2.$$

This inequality and the inequality $2xy \leq x^2/a + y^2 a$ applied to $a = 8/\delta_J$ give

(16)
$$2\lambda\|(B - \hat{B}_k)_J\|_{2,1} = 2\lambda \sum_{i \in J} \|b_i - \hat{b}_i\|_2$$
$$\leq \frac{a}{m}\lambda^2|J| + \frac{m}{a}\sum_{i \in J}\|b_i - \hat{b}_i\|_2^2$$
$$\leq \frac{a}{m}\lambda^2|J| + \frac{1}{a\delta_J}\|X\hat{B}_k - XB\|_F^2$$
$$\leq \frac{a}{m}\lambda^2|J| + \frac{2}{a\delta_J}(\hat{\Delta}_k^2 + \Delta^2)$$
$$= \frac{8\lambda^2|J|}{m\delta_J} + \frac{1}{4}(\hat{\Delta}_k^2 + \Delta^2).$$

After we combine (14), (15) and (16), we obtain

$$\frac{3}{4}\hat{\Delta}_k^2 \leq 4\lambda\|(\hat{B}_k - B)_J\|_{2,1} \leq \frac{16\lambda^2|J|}{m\delta_J} + \frac{1}{2}(\hat{\Delta}_k^2 + \Delta^2),$$

hence,

(17)
$$\hat{\Delta}_k^2 \leq 2\Delta^2 + \frac{64\lambda^2|J|}{m\delta_J} = 2\Delta^2 + 64Ck\sigma^2|J|\log(ep)\frac{\lambda_1(\Sigma)}{\delta_J}$$

for the choice $\lambda^2 = C\lambda_1(\Sigma)mk\sigma^2\log(ep)$. This concludes the first case.

Case 2. *Assume that*

(18)
$$\tfrac{5}{4}\Delta^2 + 16kd_1^2(P_{\tilde{J}}E) > 2\lambda\|(\hat{B}_k - B)_J\|_{2,1}.$$

In this case, (14) now gives

(19)
$$3\hat{\Delta}_k^2 \leq 10\Delta^2 + 64kd_1^2(P_{\tilde{J}}E).$$

By Lemma 8, we have

(20)
$$\mathbb{E}[d_1^2(P_{\tilde{J}}E)] \leq 6\sigma^2 n + 16\sigma^2 \exp(-n/2) + 6\sigma^2 \mathbb{E}\left[|\tilde{J}| + |\tilde{J}|\log\left(\frac{ep}{|\tilde{J}|}\right)\right]$$

$$\leq 6\sigma^2 n + 16\sigma^2 \exp(-n/2) + 12\sigma^2 \log(ep)(|J| + \mathbb{E}[|\hat{J}_k|]).$$

We now bound $\mathbb{E}[|\hat{J}_k|]$. Since $r(\hat{B}_k) \leq k$, we may 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 9, with V fixed at $\hat{V}_k$, $\hat{S}$ is the (globally) optimal solution to the convex problem of (25). Let x_i and $\hat{s}_i$ for $1 \leq i \leq p$ be the column vectors of X and $\hat{S}'$, respectively. Using the Karush–Kuhn–Tucker condition of $\hat{S}'$, we obtain

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

so that

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

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

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

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

$$\leq 2m\lambda_1(\Sigma)\{\hat{\Delta}_k^2 + kd_1^2(P_{\hat{J}_k}E)\}$$

using $r(\hat{V}_k) \leq k$. Taking expectations on both sides,

(22)
$$\mathbb{E}[|\hat{J}_k|] \leq \frac{2m\lambda_1(\Sigma)}{\lambda^2}\{\mathbb{E}[\hat{\Delta}_k^2] + k\mathbb{E}[d_1^2(P_{\hat{J}}E)]\}$$

$$\leq \frac{2m\lambda_1(\Sigma)}{\lambda^2}\{\mathbb{E}[\hat{\Delta}_k^2] + k\mathbb{E}[d_1^2(P_{\tilde{J}}E)]\}$$

since $\hat{J}_k \subseteq \tilde{J}$. After we combine (20) and (22), we get

$$\mathbb{E}[d_1^2(P_{\tilde{J}}E)] \leq 6\sigma^2 n + 16\sigma^2 \exp(-n/2) + 12\sigma^2 \log(ep)|J|$$

$$+ 24\sigma^2 \log(ep)\frac{m\lambda_1(\Sigma)}{\lambda^2}\{\mathbb{E}[\hat{\Delta}_k^2] + k\mathbb{E}[d_1^2(P_{\tilde{J}}E)]\}.$$

Taking $\lambda^2 = C \log(ep) mk\sigma^2 \lambda_1(\Sigma)$ with $C = 792$, we find

$$64k\mathbb{E}[d_1^2(P_{\tilde{J}}E)] \le 66k[6\sigma^2 n + 16\sigma^2 \exp(-n/2) + 12\sigma^2 \log(ep)|J|] + 2\mathbb{E}[\hat{\Delta}_k^2].$$

Inserting this bound in (19), we obtain

$$(23) \quad \mathbb{E}[\hat{\Delta}_k^2] \le 10\Delta^2 + 66k[6\sigma^2 n + 16\sigma^2 \exp(-n/2) + 12\sigma^2 \log(ep)|J|].$$

This concludes the second case. Our risk bound (8) follows directly from (17) and (23).

A.4. Proof of Theorem 4. 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 (2011) show that for $c_0 = (\sqrt{2} - 1)^2/2$,

$$\mathbb{P}\{\hat{r} \ne r\} \le \exp\{-c_0(n + q)\}.$$

Next, we decompose the risk as follows:

$$(24) \quad \begin{aligned} \mathbb{E}[\|XA - X\hat{B}^{(1)}\|_F^2] &= \mathbb{E}[\|XA - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} = r\}] \\ &\quad + \mathbb{E}[\|XA - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \ne r\}]. \end{aligned}$$

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 $\hat{r} = r(\hat{B}) \ge r(O) = 0$, we have the inequality

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

by the minimzing property of $\hat{B}^{(1)}$. Using Pythagoras, $\|PY - X\hat{B}^{(1)}\|_F^2 \le \|PY\|_F^2$ for P the projection matrix on the column space of X. Consequently,

$$\begin{aligned} \mathbb{E}[\|XA &- X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \ne r\}] \\ &\le 2\mathbb{E}[\{\|PY - XA\|_F^2 + \|PY - X\hat{B}^{(1)}\|_F^2\}I\{\hat{r} \ne r\}] \\ &\le 2\mathbb{E}[\{\|PE\|_F^2 + \|PY\|_F^2\}I\{\hat{r} \ne r\}] \\ &\le 6\mathbb{E}[\|PE\|_F^2 I\{\hat{r} \ne r\}] + 4\|XA\|_F^2 \mathbb{P}\{\hat{r} \ne r\}. \end{aligned}$$

Since $\|PE\|_F^2/\sigma^2$ has a Chi-square distribution with qn degrees of freedom [see Lemma 3 in Bunea, She and Wegkamp (2011)], we have

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

We obtain, using the Cauchy–Schwarz inequality,

$$\mathbb{E}[\|PE\|_F^2 I\{\hat{r} \ne r\}] \le \sqrt{2}qn\sigma^2 \exp\{-c_0(n + q)/2\}$$

and so

$$\mathbb{E}[\|XA - X\hat{B}^{(1)}\|_F^2 I\{\hat{r} \neq r\}] \leq 4\|XA\|_F^2 \exp\{-c_0(n+q)\}$$
$$+ 6\sqrt{2}qn\sigma^2 \exp\{-c_0(n+q)/2\}.$$

The second term on the right is clearly bounded. For the first term on the right, we invoke condition $\mathfrak{C}_2$ on $\|XA\|_F^2$. This completes our proof.

A.5. Proof of Theorem 5. In this proof, $C_1, \ldots, C_7$ are numerical, positive and finite constants. Theorem 2 of Section 2, applied to the random matrices $\hat{B}_{\lambda,k}$, yields

$$\mathbb{E}[\|X\hat{B}^{(2)} - XA\|_F^2]$$
$$\leq \min_{k,\lambda} \mathbb{E}[10\|X\hat{B}_{\lambda,k} - XA\|_F^2 + 8\,\mathrm{pen}(\hat{B}_{\lambda,k})] + 768n\sigma^2 \exp(-n/2)$$
$$\leq C_1 \min_{\lambda} \mathbb{E}[\|X\hat{B}_{\lambda,r} - XA\|_F^2 + \mathrm{pen}(\hat{B}_{\lambda,r}) + n\sigma^2].$$

(Here we used $c = 12$ in the penalty term.) Since Σ satisfies $\mathfrak{A}(J(A), \delta_{J(A)})$, and we assume that $\lambda_1(\Sigma)/\delta_{J(A)} \leq C_3$, Theorem 3 yields, for each global solution $\hat{B}_{\lambda,r}$,

$$\mathbb{E}[\|X\hat{B}_{\lambda,r} - XA\|_F^2] \leq C_2\{n + |J(A)|\log(p)\}r\sigma^2$$

for $\lambda^2 = C\sigma^2\lambda_1(\Sigma)km\log(ep)$ with C large enough. It remains to bound the expected penalty term $\mathbb{E}[\mathrm{pen}(\hat{B}_{\lambda,r})]$. Since

$$\mathbb{E}[\mathrm{pen}(\hat{B}_{\lambda,r})] \leq C_4(n + \log(p)\mathbb{E}[|J(\hat{B}_{\lambda,r})|]),$$

we need to bound $\mathbb{E}[|J(\hat{B}_{\lambda,r})|]$. We write $\hat{J}_r = J(\hat{B}_{\lambda,r})$. From (22) in the proof of Theorem 3, we have

$$\mathbb{E}[|\hat{J}_r|] \leq \frac{2m\lambda_1(\Sigma)}{\lambda^2}\{\mathbb{E}[\|X\hat{B}_{\lambda,r} - XA\|_F^2] + r\mathbb{E}[d_1^2(P_{\hat{J}_r}E)]\},$$

while Lemma 8 gives

$$\mathbb{E}[d_1^2(P_{\hat{J}_r}E)] \leq 6n\sigma^2 + 16\sigma^2 \exp(-n/2) + 12\sigma^2 \log(ep)\mathbb{E}[|\hat{J}|].$$

Therefore, taking $\lambda^2 = C\sigma^2\lambda_1(\Sigma)km\log(ep)$ with C large enough, we obtain from the previous three displays

$$\mathbb{E}[|\hat{J}_r|] \leq C_5\{\mathbb{E}[\|X\hat{B}_{\lambda,r} - XA\|_F^2]/r + n\sigma^2/\log(p)\}$$
$$\leq C_6\sigma^2\{n + |J(A)|\log(p)\}.$$

We now can conclude that

$$\mathbb{E}[\|X\hat{B}^{(2)} - XA\|_F^2] \leq C_7\sigma^2 r\{n + \log(p)|J(A)|\},$$

and the proof is complete.

A.6. Proof of Theorem 6. In this proof we drop the subscripts λ, k in the (S, V) iterates.

PROOF OF PART (ii). The proof of this part of Theorem 6 will follow from the global convergence theorem (GCT) of Zangwill and Mond (1969). For completeness, we state this theorem below, then we verify that its conditions hold in Lemmas 9, 10 and 11.

THE GLOBAL CONVERGENCE THEOREM [Luenberger and Ye (2008)]. *Let A be a map describing an algorithm on $\mathcal{X}$ 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 $\Gamma \subset \mathcal{X}$ be given, and suppose:*

(1) All points x_k are contained in a compact set $S \subset \mathcal{X}$;
(2) There exists a continuous function Z on $\mathcal{X}$ 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 Γ.

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 $\mathscr{A}$. 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:

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

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

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 $\mathscr{A}$ 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

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

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 9 shows the algorithm converges globally for any initial starting point.

LEMMA 9. *For any $j \geq 0$, $\mathrm{rank}(B^{(j)}) \leq k$ and $F(B^{(j)}) \geq F(B^{(j+1)})$.*

PROOF. We write $F(S, V) = \tfrac{1}{2} \|YV - XS\|_F^2 + \lambda \|S\|_{2,1} + \tfrac{1}{2} \mathrm{tr}(Y(I - VV')Y')$. Given V, (10) reduces to the following optimization problem after

vectorization:

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

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 $F(S,V) = -\operatorname{tr}(Y'XSV') + \|Y\|_F^2/2 + \|XS\|_F^2/2 + \lambda\|S\|_{2,1}$, we see the optimization problem is equivalent to

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

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

(27) $$\hat{V} = U_w V_w'$$

achieves the upper bound $\sum_i d_i(W)$. This globally optimal solution to (26) is the one used in Algorithm $\mathscr{A}$. 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, $F(B_{k,\lambda}^{(j)}) \geq F(B_{k,\lambda}^{(j+1)})$ during each iteration. $\square$

LEMMA 10. *Suppose $\lambda > 0$. Then $B^{(j)}$, $S^{(j)}$, $V^{(j)}$ in Algorithm $\mathscr{A}$ are uniformly bounded in j.*

PROOF.   From Lemma 9, $\|S^{(j)}\|_{2,1} \leq F(S^{(j)}, V^{(j)})/\lambda \leq F(S^{(1)}, V^{(1)})/\lambda \leq F(0, V^{(0)})/\lambda \leq \|Y\|_F^2/(2\lambda)$. Therefore, $\|S^{(j)}\|$ must be uniformly bounded. $\square$

LEMMA 11. *Suppose $\lambda > 0$. The map $\mathcal{M}$ introduced for describing Algorithm $\mathscr{A}$ is a closed point-to-set map on Ω.*

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 $\mathbb{R}^{n \times k}$.
- $\mathcal{M}^S(\omega) \neq \varnothing, \mathcal{M}^V(\omega) \neq \varnothing, \forall \omega \in \Omega$, seen from the proof of Lemma 9.

First, we prove that $\mathcal{M}^S$ is closed on Ω. It suffices to show the point-to-set map $\mathcal{M}^s(x) = \{y \in \mathbb{R}^{p \times k} : F(y,x) \leq \min_{\tilde{y} \in \mathbb{R}^{p \times k}} F(\tilde{y}, x)\}$ is closed at any $x \in \mathbb{O}^{n \times k}$. Let $x_j \to x^*, y_j \in \mathcal{M}^s(x_j)$ and $y_j \to y^*$, 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 $y^* \in \mathbb{R}^{p \times k}$, $F(y^*, x^*) > \min_{y \in \mathbb{R}^{p \times k}} F(y, x^*) =: L$. There exists some $\varepsilon_0 > 0$ such that $F(y^*, x^*) > L + \varepsilon_0$. Let $\tilde{y} \in \mathcal{M}^s(x^*)$. Then $F(\tilde{y}, x^*) = L$. Since F is continuous at $(\tilde{y}, x^*)$, $\lim_{j \to \infty} F(\tilde{y}, x_j) = L$. For j large enough, $|F(\tilde{y}, x_j) - L| \leq \varepsilon_0/2$, from which it follows that $F(y_j, x_j) \leq F(\tilde{y}, x_j) \leq L + \varepsilon_0/2$ and $F(y^*, x^*) \leq L + \varepsilon_0/2$.

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 10, 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), page 205], $\mathcal{M}$ is closed on Ω. $\square$

If we set Z in the general statement of the global convergence theorem to be our continuous criterion function F, and if we take Γ to be the set of local of minima of F, Lemma 9 and the assumption in part (ii) of our Theorem 7 guarantee that (2) of GCT holds. Lemmas 10 and 11 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 $\mathscr{A}$ converges to a stationary point of F. We begin by defining a stationary point in this context. Recall that we view (10) 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 (1986)]. This Riemannian gradient, denoted by $\nabla_V F(S, V)$, can be explicitly computed: $\nabla_V F(S, V) = \mathcal{P}_V(-W) = -W + VV'W/2 + VW'V/2$ 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), page 62], a necessary condition for F to have a local minimum at (S^*, V^*) is that this point is stationary, that is,

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

For future use, note that $\nabla_V F(S, V)$ is continuous on Ω. 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 and Mond (1969), Bertsekas (1999) and Luenberger and Ye (2008) for details]. Therefore, GCD cannot be directly applied.

From Lemma 9, $F(S^{(j)}, V^{(j)})$ must converge. Denote the limit by L^*. Let $(S^{(j_l)}, V^{(j_l)})$ $(l = 1, 2, \ldots)$ be a subsequence of $(S^{(j)}, V^{(j)})$ which converges to (S^*, V^*) as $l \to \infty$. Then $L^* = \lim_{l \to \infty} F(S^{(j_l)}, V^{(j_l)}) = F(S^*, V^*)$. We assume, without loss of generality, $(S^{(j_l+1)}, V^{(j_l+1)})$ also converges (Lemma 10), and denote the limit by $(\bar{S}, \bar{V})$. We have

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

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 11, however, states that $(\bar{S}, \bar{V}) \in \mathcal{M}(S^*, V^*)$. That is, $L^* = F(\bar{S}, \bar{V}) \leq F(\bar{S}, V^*) < F(S^*, V^*) = L^*$, 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.

A.7. Proof of Theorem 7. We use the same notation system as in Appendix A.6. 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 $F(S^*, V) \geq F(S^*, V^*)$, $\forall V \in \mathbb{O}^{n \times k}$ [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 $\mathscr{A}'$. For simplicity, set $K = 1$ in (11) and redefine the operator $T_V : \mathbb{R}^{p \times k} \to \mathbb{R}^{p \times k}$ by

$$(30) \qquad T_V \circ S = \vec{\Theta}(X'YV + (I - X'X)S; \lambda) \qquad \forall S \in \mathbb{R}^{p \times k}.$$

Let T_V^α ($\alpha \in \mathbb{N}$) be the composition of α T_V's. Define point-to-set maps $\mathcal{M}^S(S, V) = \{(\bar{S}, \bar{V}) \in \Omega : \bar{S} \in \{T_V \circ S, T_V^2 \circ S, \ldots, T_V^{M_{\text{iter}}} \circ S\}, \bar{V} = V\}$, and $\mathcal{M}^V(S, V) = \{(\bar{S}, \bar{V}) \in \Omega : \inf_{\tilde{V} \in \mathbb{O}^{n \times k}} F(S, \tilde{V}) \geq F(S, \bar{V}), \bar{S} = S\}$. Then $\mathcal{M} = \mathcal{M}^V \mathcal{M}^S$ characterizes $\mathscr{A}'$. When updating S at step (a), the algorithm allows one to perform T any times (denoted by α_j) provided α_j does not go beyond $M_{\text{iter}} \in \mathbb{N}$ that is prespecified before running $\mathscr{A}'$.

LEMMA 12. *Given any* $V \in \mathbb{O}^{n \times k}$ *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 (2012) to the vectorized problem (25). Note that $2 - \|X \otimes I\|_2 = 2 - \|X\|_2 \geq 1$. The proof details are omitted. □

Choose $(\bar{S}, V) \in \mathcal{M}^S(S, V)$, using the triangle inequality we know $F(S, V) - F(\bar{S}, V) \geq \frac{1}{2}\|\bar{S} - S\|_F^2$.

LEMMA 13. *Suppose* $\lambda > 0$. *Then* $B^{(j)}$, $S^{(j)}$, $V^{(j)}$ *in* $\mathscr{A}'$ *are uniformly bounded in* j.

The proof is similar to the proof of Lemma 10 and therefore omitted.

LEMMA 14. *Suppose* $\lambda > 0$. *The* $\mathcal{M}$ *for describing* $\mathscr{A}'$ *is a closed point-to-set map on* Ω.

PROOF. Similar to the proof of Lemma 11, we prove that the point-to-set map $\mathcal{M}^s(S, V) = \{T_V \circ S, T_V^2 \circ S, \ldots, T_V^{M_{\text{iter}}} \circ S\}$ is closed at any $(S, V) \in \Omega$. Then M^S and thus $\mathcal{M}$ are closed on Ω.

Let $(S_j, V_j) \to (S^*, V^*)$, $\tilde{S}_j \in \mathcal{M}^s(S_j, V_j)$ and $\tilde{S}_j \to \tilde{S}^*$, with (S_j, V_j), $(S^*, V^*) \in \Omega$ 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. $\quad\square$

Now we prove Theorem 7. Following the lines of the proof of part (i) of Theorem 6, for any accumulation point (S^*, V^*) of $(S^{(j)}, V^{(j)})$, $(S^*, V^*) \in \Omega$ and there exists $(\bar{S}, \bar{V}) \in \mathcal{M}(S^*, V^*)$ with $F(\bar{S}, \bar{V}) = F(S^*, V^*)$. Since $F(\bar{S}, \bar{V}) \le F(\bar{S}, V^*) \le F(S^*, V^*)$, $F(\bar{S}, V^*) = F(S^*, V^*)$. It follows from the comment after Lemma 12 that $\bar{S} = S^*$. 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 $\alpha \le \alpha_0$, and, in particular, $F(T_{V^*} \circ S^*, V^*) = F(S^*, V^*)$. Applying Lemma 12 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 $\bar{S} = S^*$, we have $(S^*, \bar{V}) \in \mathcal{M}^V(S^*, V^*)$. $\bar{V}$ is a (global) minimizer of $F(S^*, V)$ given S^*. But $F(S^*, \bar{V}) = F(\bar{S}, \bar{V}) = F(S^*, V^*)$, 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

ALDRIN, M. (1996). Moderate projection pursuit regression for multivariate response data. *Comput. Statist. Data Anal.* **21** 501–531. MR1395234

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22** 327–351. MR0042664

BERTSEKAS, D. (1999). *Nonlinear Programming*. Athena Scientific, Nashua, NH.

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOOTHBY, W. M. (1986). *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed. *Pure and Applied Mathematics* **120**. Academic Press, Orlando, FL. MR0861409

BRILLINGER, D. R. (1981). *Time Series: Data Analysis and Theory*, 2nd ed. Holden-Day, Oakland, CA. MR0595684

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

BUNEA, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat.* **2** 1153–1194. MR2461898

BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355

BUNEA, F., SHE, Y., OMBAO, H., GONGVATANA, A., DEVLIN, K. and COHEN, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *NeuroImage* **55** 1519–1527.

CANDÈS, E. J. and PLAN, Y. (2010). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inform. Theory* **57** 2342–2359.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

GABAY, D. (1982). Minimizing a differentiable function over a differential manifold. *J. Optim. Theory Appl.* **37** 177–219. MR0663521

GIRAUD, C. (2011). Low rank multivariate regression. *Electron. J. Stat.* **5** 775–799. MR2824816

IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York. MR2445017

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

LUENBERGER, D. G. and YE, Y. (2008). *Linear and Nonlinear Programming*, 3rd ed. *International Series in Operations Research & Management Science* **116**. Springer, New York. MR2423726

NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. MR2816348

REINSEL, G. C. and VELU, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications. Lecture Notes in Statistics* **136**. Springer, New York. MR1719704

ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comput. Statist. Data Anal.* **56** 2976–2990.

SHIMIZU, K., ISHIZUKA, Y. and BARD, J. F. (1997). *Nondifferentiable and Two-Level Mathematical Programming*. Kluwer Academic, Boston, MA. MR1429865

TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

WEI, F. and HUANG, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16** 1369–1384. MR2759183

WU, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

YUAN, M., EKICI, A., LU, Z. and MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 329–346. MR2323756

ZANGWILL, W. I. and MOND, B. (1969). *Nonlinear Programming: A Unified Approach*. Prentice Hall International, Englewood Cliffs, NJ. MR0359816

F. BUNEA
DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
1198 COMSTOCK HALL
ITHACA, NEW YORK 14853-3801
USA
E-MAIL: fb238@cornell.edu

Y. SHE
DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
117 N. WOODWARD AVE
TALLAHASSEE, FLORIDA 32306-4330
USA
E-MAIL: yshe@stat.fsu.edu

M. H. WEGKAMP
DEPARTMENT OF MATHEMATICS
AND
DEPARTMENT OF STATISTICAL SCIENCE
CORNELL UNIVERSITY
1194 COMSTOCK HALL
ITHACA, NEW YORK 14853-3801
USA
E-MAIL: marten.wegkamp@cornell.edu