# Joint Variable and Rank Selection for Parsimonious Estimation of High-Dimensional Matrices

**Marten Wegkamp**

Department of Mathematics
Department of Statistical Science
**Cornell University**

**June 2012**

Talk based on:

- *Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices.*
  (with Florentina Bunea and Yiyuan She)
  Annals of Statistics 39(2), 1282-1309 (2011).
- *Joint Variable and Rank Selection for Parsimonious Estimation of High-Dimensional Matrices*
  (with Florentina Bunea and Yiyuan She)

# Multivariate Response Regression Model

Observations $(X_1, Y_1), \ldots, (X_m, Y_m) \in \mathbb{R}^p \times \mathbb{R}^n$ related via regression model

$$Y = XA + E$$

- $X$: $m \times p$ design matrix of rank $q$
- $A$: $p \times n$ matrix of unknown coefficients
- $E$: $m \times n$ matrix of independent $N(0, \sigma^2)$ errors $E_{ij}$

- Standard least squares estimation under no constraints = regressing each response on the predictors separately.
- It completely ignores the multivariate nature of the possibly correlated responses.

Problem: We need to estimate $A$, that is, $nq$ parameters!

Solution: Impose matrix sparsity!

Let $r$ be the rank of $A$ and $|J|$ be the number of non-zero rows of $A$. Number of free parameters (in SVD of $A$) is in fact

$$r(n + |J| - r).$$

Of course, $r$ and $J$ are unknown.

## Solutions:

- (variable selection)
  GLASSO: *Yuan and Lin ( 2006); Lounici, Pontil, Tsybakov and van de Geer (2011)*
- (rank selection)
  RSC: *Bunea, She, Wegkamp (2011), Giraud (2011), Klopp (2011)*
  NNP: *Candès and Plan (2011), Rhode and Tsybakov (2011), Negahban and Wainwright (2011), Bunea, She, Wegkamp (2011)*
- (joint rank and row selection)
  JRRS: *Bunea, She, Wegkamp (2011)*.

Number of free parameters: $r(n + |J| - r)$.

Note $m, p, n, q, r, |J|$ satisfy $q \leq m \wedge p$, $r \leq n \wedge |J|$, $|J| \leq q$.

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

Improvement possible for $n < q$. Since $(|J| + n)r \leq (q + n)r$ and $(n + |J|)r \leq 2(n \vee |J|)(n \wedge |J|) \leq 2|J|n$, JRRS often wins.

# Part I: Rank sparsity

# A historical perspective and existing results

Estimation under the constraint $\text{rank}(A) = r$, with $r$ *known.*

- Anderson (1951, 1999, 2002)
- Robinson (1973, 1974)
- Izenman (1975; 2008)
- Rao (1979)
- Reinsel and Velu (1998)

All theoretical results (distribution of the reduced rank estimates and rank selection procedures) are asymptotic, $m \to \infty$, everything else fixed.

# A finite sample approach to dimension reduction

We derive reduced rank estimates $\widehat{A}$, without prior specification of the rank.

- We propose a computationally efficient method that can handle matrices of large dimensions.
- We provide a finite sample analysis of the resulting estimates.
- Our analysis is valid **for any** $m$, $n$, $p$ and $r$.

# Methodology

We propose to estimate $A$ by the penalized least squares estimator

$$\begin{aligned}\widehat{A} &= \arg\min_{B}\{\|Y - XB\|_F^2 + \mu \cdot r(B)\} \\ &= \arg\min_{B}\{\|PY - XB\|_F^2 + \mu \cdot r(B)\}\end{aligned}$$

for projection $P$ on $X$.

Set $\widehat{k} = r(\widehat{A})$ and let $\widehat{B}_k$ be the restricted LSE of rank $k$. Then

$$\begin{aligned} \|Y - X\widehat{A}\|_F^2 + \mu \cdot \widehat{k} &= \min_B \{\|Y - XB\|_F^2 + \mu \cdot r(B)\} \\ &= \min_k \{\|Y - X\widehat{B}_k\|_F^2 + \mu \cdot k\} \end{aligned}$$

# Closed form solutions

Our first result states that $\widehat{A}$, $X\widehat{A}$ and $\widehat{k} = r(\widehat{A})$ have closed form solutions and can be efficiently computed based on the SVD of $PY$.

**Proposition**

- $\widehat{k}$ is the number of singular values of $PY$ that exceed $\sqrt{\mu}$
- $X\widehat{A} = \sum_{j \leq \widehat{k}} d_j u_j v_j'$
- $\widehat{A}$ is the rank restricted LSE (of rank $\widehat{k}$)

# Consistent Effective Rank Estimation

**Theorem**

Suppose that there exists an index $s \le r$ such that

$$d_s(XA) > (1+\delta)\sqrt{\mu}$$

and

$$d_{s+1}(XA) < (1-\delta)\sqrt{\mu},$$

for some $\delta \in (0,1]$. Then we have

$$\mathbb{P}\left\{\widehat{k} = s\right\} \ge 1 - \mathbb{P}\left\{d_1(PE) \ge \delta\sqrt{\mu}\right\}.$$

- We can consistently estimate the index $s$ provided we use a large enough value for $\mu$ to guarantee that the probability of the event $\left\{d_1(PE) \leq \delta\sqrt{\mu}\right\}$ approaches one.
- We call $s$ the *effective rank* of $A$ relative to $\mu$, and denote it by $r_e = r_e(\mu)$.
- We can only hope to recover those singular values of the signal $XA$ that are above the noise level $d_1(PE)$. Their number, $r_e$, will be the target rank of the approximation of the mean response, and can be much smaller than $r = r(A)$.
- The largest singular value $d_1(PE)$ is our relevant indicator of the strength of the noise.

**Lemma**

Let $q = r(X)$ and assume that $E_{ij}$ are independent $N(0, \sigma^2)$ random variables. Then

$$\mathbb{E}\left[d_1(PE)\right] \leq \sigma\left(\sqrt{n} + \sqrt{q}\right)$$

and, for all $t > 0$,

$$\mathbb{P}\left\{d_1(PE) \geq \mathbb{E}[d_1(PE)] + \sigma t\right\} \leq \exp\left(-t^2/2\right).$$

In view of this result, we take

$$\mu = 2\sigma^2(n+q)$$

as our measure of the noise level.

Summarizing,

**Corollary**

If $d_r(XA) > 2\sqrt{\mu}$, then $\mathbb{P}\{\widehat{k} = r\} \to 1$ as $q + n \to \infty$.

# Risk Bounds for the Restricted Rank LSE

**Theorem**

Let $\widehat{B}_k$ be the restricted LSE of rank $k$. For every $k$ we have

$$\|X\widehat{B}_k - XA\|_F^2 \leq 3\left[\sum_{j>k} d_j^2(XA) + 4k d_1^2(PE)\right]$$

with probability one.

# Risk Bounds for the Restricted Rank LSE

- We bound the error $\|X\widehat{B}_k - XA\|_F^2$ by an approximation error, $\sum_{j>k} d_j^2(XA)$, and a stochastic term, $kd_1^2(PE)$.
- The approximation error is decreasing in $k$ and vanishes for $k > r(XA)$.
- The stochastic term can be bounded by $C\sigma^2 k(n+q)$ with large probability, and is increasing in $k$.
- $k(n+q)$ is essentially the number of free parameters of the restricted rank problem as the parameter space consists of all $p \times n$ matrices $B$ of rank $k$ and each matrix has $k(n+q-k)$ free parameters.
- The obtained risk bound is the squared bias plus the dimension of the parameter space.

# Risk Bound for the RSC Estimator

**Theorem**

We have, for any $\mu$,

$$\mathbb{P}\left[\|X\widehat{A} - XA\|_F^2 \leq 3\left\{\|XB - XA\|_F^2 + \mu r(B)\right\}\right]$$
$$\geq \quad 1 - \mathbb{P}\left[2d_1(PE) > \sqrt{\mu}\right],$$

for all $p \times n$ matrices $B$.

# Risk Bound for the RSC Estimator

**Theorem**

In particular, we have, for $\mu = C_0\sigma^2(q+n)$ and some $C_0 > 1$,

$$\mathbb{E}\left[\|X\widehat{A} - XA\|_F^2\right] \leq C \min_k \left\{\sum_{j>k} d_j^2(XA) + \sigma^2(q+n)k\right\}.$$

# Remarks

- RSC achieves optimal bias-variance trade-off.
- RSC is minimax adaptive.
- Minimizer of $\sum_{j>k} d_j^2(XA) + \mu k$ is effective rank $r_e$.
- RSC adapts to $r_e$.
- The smaller $r$, the smaller the prediction error.
- Bounds valid for all $m, n, p, q, r$.

# Part II: Joint Rank and Row Sparsity

Minimize

$$\|Y - XB\|_F^2 + c\sigma^2 r(B)\left\{2n + 2|J(B)| + |J(B)|\log\left(\frac{p}{2|J(B)|}\right)\right\}$$

over *all* $p \times n$ matrices $B$. Here $c > 3$ is a numerical constant.

**Theorem**

For any $c > 3$,

$$\begin{aligned}\mathbb{E}\left[\|XA - X\widehat{B}\|_F^2\right] &\lesssim \inf_B \left[\|XA - XB\|_F^2 + \text{pen}(B)\right] \\ &\lesssim \sigma^2 r(A)\left\{n + |J(A)|\log\left(\frac{p}{|J(A)|}\right\}.\end{aligned}$$

## Remarks

- $\widehat{B}$ adapts to the unknown row and rank sparsity of $A$

## Two-step procedures

- First select rank, then rows.
- First select rows, then rank.

# Method 1

## Method 1

- Use RSC to select

$$\widehat{r} = \sum_{k} 1\{d_k(PY) \geq \sigma(\sqrt{2n} + \sqrt{2q})\}$$

- Use RCGL $\widehat{B}_k$ with $k = \widehat{r}$ to obtain final estimator

# Rank Constrained Group Lasso

$$\widehat{B}_k = \underset{rank(B) \leq k}{\arg\min} \left\{ \| Y - XB \|_F^2 + 2\lambda \| B \|_{2,1} \right\}.$$

with $\lambda = C\sigma\sqrt{mk}\sqrt{\lambda_1(X'X/m)}$

- $k = n$: no rank restriction (GLASSO)
- $\lambda = 0$: reduced-rank regression

# Assumption

## Assumption $\mathfrak{A}$ on Gram matrix

Set $\Sigma = X'X/m$. There exists a set $I \subseteq \{1, \ldots, p\}$ and $\delta_I > 0$ such that

$$\mathrm{tr}(M'\Sigma M) \geq \delta_I \sum_{j \in I} \|m_j\|_2^2$$

for all $p \times n$ matrices $M$ with rows $m_j$ satisfying

$$\sum_{j \in I} \|m_j\|_2 \geq 2 \sum_{j \notin I} \|m_j\|_2.$$

# Row-sparse adaptive

**Theorem**

Let $\widehat{B}_k$ be a global minimizer. Then, for any $p \times n$ matrix $B$ with $r(B) \leq k$ and $|J(B)|$ non-zero rows,

$$\mathbb{E}\left[\|X\widehat{B}_k - XA\|_F^2\right]$$
$$\lesssim \|XB - XA\|_F^2 + k\sigma^2\left\{n + \left(1 + \frac{\lambda_1(\Sigma)}{\delta_{J(B)}}\right)|J(B)|\log(p)\right\},$$

provided $\Sigma$ satisfies Assumption $\mathfrak{A}(J(B), \delta_{J(B)})$.

- If the generalized condition number $\lambda_1(\Sigma)/\delta_{J(B)}$ is bounded, then, within the class of row sparse matrices of fixed rank $k$, the RCGL estimator is row-sparsity adaptive.
- Moreover, if the rank $r$ of $A$ is known, then RCGL achieves the desired rate of convergence in row and rank sparse models.
- GLASSO minimizes criterion over *all* $p \times n$ matrices $B$.
  Optimal choice $\lambda = 2\sqrt{2}\sigma\sqrt{mn}\left(1 + \frac{A \log p}{n}\right)^{1/2}$, see Lounici et al (2011).
  Our choice replaces $n$ by $k$: we minimize over all rank-$k$ matrices!

### Condition $\mathfrak{C}$ on signal

$\mathfrak{C}1$ $d_r(XA) > 2\sqrt{2}\sigma(\sqrt{n} + \sqrt{q})$

$\mathfrak{C}2$ $\log(\|XA\|_F) \leq (\sqrt{2} - 1)^2(n + q)/4.$

**Theorem**

Let $\Sigma$ satisfy $\mathfrak{A}(J(A), \delta_{J(A)})$, let $\lambda_1(\Sigma)/\delta_J$ be bounded, and let $\mathfrak{C}$ hold. Then the two-step JRRS estimator $\widehat{B}^{(1)}$ satisfies

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

Conclusion:
$\widehat{B}^{(1)}$ is *row and rank* adaptive.

# Method 2

## Method 2

- Minimize

$$\|Y - XB\|_F^2 + 2\lambda \|B\|_{2,1}$$

with

$$\lambda = 2\sigma\sqrt{mn}\sqrt{1 + \frac{A \log p}{n}}$$

- Set

$$\widehat{J} = \left\{ j : n^{-1/2} \|\widehat{B}_j\| > c m^{-1/2} [1 + A \log p/n]^{1/2} \right\}$$

- Run RSC on restricted dimensions: $X_{\widehat{J}}$.

This works, provided

$$|\Sigma_{ij}| \leq \frac{1}{7\alpha|J|}$$

and

$$n^{-1/2}\|A_j\| \geq Cm^{-1/2}\left[1 + \frac{A\log p}{n}\right]^{1/2}.$$

See Lounici et al (2011) for consistency of $\widehat{J}$.

# Simulation setup

- $X$ has i.i.d. rows $X_i$ from MVN$(\mathbf{0}, \Sigma)$, with $\Sigma_{jk} = \rho^{|j-k|}$, $\rho > 0$, $1 \leq j, k \leq p$.
- $$A = \begin{bmatrix} A_1 \\ O \end{bmatrix} = \begin{bmatrix} bB_0 B_1 \\ O \end{bmatrix},$$
  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. $N(0, 1)$.
- $E_{ij}$ are iid $N(0, 1)$.

We report two settings:

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

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

We tested four methods: RSC, GLASSO, method 1 and method 2.

Table: $p$ large

| | **MSE** | $\lvert\widehat{J}\rvert$ | $\widehat{R}$ | 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** | 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** | **402** | 14 | 2 | 41% | 7% |

Table: $m$ large

| | **MSE** | $\lvert\widehat{J}\rvert$ | $\widehat{R}$ | 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** | **8.9** | 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** | **8.1** | 15 | 5 | 0% | 0% |

# Conclusions

- GLASSO often severely misses some true features in the large-$p$ setup as seen from its high M numbers.
- RSC achieved good rank recovery. The drawback is that this dimension reduction requires using all $p$ variables and thus hurts interpretability.
- Clearly both GLASSO and RSC are inferior to the two JRRS methods.
- Method 1 (RSC→RCGL) dominates all other methods. Its MSE results are impressive. While it may not give exactly $|\widehat{J}| = |J| = 15$, its M numbers indicate that we did not miss many true features.

Thanks!

# Efficient Computation of $\widehat{B}_k$ (Reinsel and Velu, 1998).

Let $M = X'X$ be the Gram matrix, and let $P = XM^{-}X'$.

1. Compute the eigenvectors $V = [v_1, v_2, \cdots, v_n]$, corresponding to the ordered eigenvalues arranged from largest to smallest, of the symmetric matrix $Y'PY$.
2. Compute $\widehat{B} = M^{-}X'Y$.
   Construct $W = \widehat{B}V$ and $G = V'$.
   Form $W_k = W[\,,1:k]$ and $G_k = G[1:k,\,]$.
3. Compute the final estimator $\widehat{B}_k = W_k G_k$.

## Algorithm

- Given $1 \leq k \leq m \wedge p \wedge n$, $\lambda \geq 0$, $V_{k,\lambda}^{(0)} \in \mathbb{O}^{n\times k}$.
- $j \leftarrow 0$, converged $\leftarrow$ FALSE
  WHILE not converged
  - $S_{k,\lambda}^{(j+1)} \leftarrow \arg\min_{S\in\mathbb{R}^{p\times k}} \frac{1}{2}\|YV_{k,\lambda}^{(j)} - XS\|_F^2 + \lambda\|S\|_{2,1}$.
  - Let $W \leftarrow Y'XS_{k,\lambda}^{(j+1)} \in \mathbb{R}^{n\times k}$ and perform SVD: $W = U_w D_w V_w'$ with $D_w$ diagonal.
  - $V_{k,\lambda}^{(j+1)} \leftarrow U_w V_w'$
  - $B_{k,\lambda}^{(j+1)} \leftarrow S_{k,\lambda}^{(j+1)}(V_{k,\lambda}^{(j+1)})'$
  - converged $\leftarrow \|B_{k,\lambda}^{(j+1)} - B_{k,\lambda}^{(j)}\|_\infty < \varepsilon$
  - $j \leftarrow j+1$

  ENDWHILE
- Deliver $\widehat{B}_{k,\lambda} = B_{k,\lambda}^{(j+1)}$, $\widehat{S}_{k,\lambda} = S_{k,\lambda}^{(j+1)}$, $\widehat{V}_{k,\lambda} = V_{k,\lambda}^{(j+1)}$.