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

Florentina Bunea
Department of Statistical Science
Cornell University

High-dimensional Problems in Statistics Workshop
ETH, September 2011
1 Framework and motivation

2 Joint Rank and Row Selection JRRS Methods
   - The construction of the one-step JRRS estimator
   - Row and rank sparsity oracle inequalities via one-step JRRS
   - One-step JRRS to select the best estimator from a finite list

3 Two-step JRRS estimators
   - Rank Constrained Group Lasso RCGL
   - Adaptive RCGL for joint row and rank selection
   - Row and rank sparsity oracle inequalities via two-step JRRS

4 Numerical performance and examples

5 Summary
A rank and row sparse model

- Model: \( Y = XA + E; \ E \) noise matrix.
- Data: \( m \times n \) matrix \( Y \) and \( m \times p \) matrix \( X \).
- Target: \( p \times n \) matrix \( A \) \( \leftrightarrow \) \( pn \) unknown parameters
- Rank of \( A \) is \( r \leq n \& p \). Nbr of non-zero rows of \( A \) is \( |J| \leq p \).
- Row and Rank Sparse Target \( \leftrightarrow \) \( r(|J| + n - r) \) free param.
- Full rank + all rows + large \( n \) and \( p \) = Hopeless, if \( m \) small.
  Low rank + Small \( |J| \) = HOPE, if \( m \) small.
- Estimate \( A \) under \textit{joint rank and row} constraints.
Why rank and row sparse $Y = XA + E$?

- **Multivariate response regression**
  
  Measure $n$ response variables for $m$ subjects: $Y_i \in \mathbb{R}^n$, $1 \leq i \leq m$.
  
  Measure $p$ predictor variables for $m$ subjects: $X_i \in \mathbb{R}^p$, $1 \leq i \leq m$.
  
  No (rank / row ) constraints on $A \iff n$ separate univ.
  
  Zero rows in $A \iff$ Not all predictors in the model.
  
  Low rank of $A \iff$ Only few orthogonal scores relevant.

**Goal:** Estimation tailored to row and rank sparsity

Use only a subset of the predictors to construct few scores, with high predictive power, under JOINT rank and row restrictions on $A$. 
Why row and rank sparse $Y = XA + E$? Contd.

- Supervised **row and rank sparse** PCA.
- Provides framework for **row and rank sparse** PCA and CCA.
- **Building block** in functional data analysis (with predictors).
  
  $Y =$ matrix of discretized trajectories for $n$ subjects;
  $X =$ matrix of basis functions evaluated at discrete data points 
  + possibly other predictors of interest.

- **Building block** in multiple time series analysis.
  (Macro-economics and forecasting)

  $Y =$ matrix of $n$ time series observed over $m$ time periods 
  ($n$ types of interest rates)
  $X =$ $Y$ in the past + other predictive time series 
  (other potentially connected macro-economic factors).
A historical perspective on sparse $Y = XA + E$

Rank Sparse Models

- Reduced-Rank Regression: $Y = XA + E$, rank $(A) = k = \text{known}$. Asymptotic results $m \to \infty$: Anderson (1951, 1999, 2002); Rao (1979); Reinsel and Velu (1998); Izenman (1975; 2008).

- Low rank approximations: $Y = XA + E$, rank $(A) = r = \text{unknown}$. Adaptive estimation + Finite sample theoretical analysis, valid for any $m, n, p$ and any $r$.

*Rank Selection Criterion (RSC)*: Bunea, She and Wegkamp (2011).

Row-Sparse Models

- Predictor $X_j$ not in the model $\iff$ The j-th row of $A$ is zero.
- Individual variable selection in multivariate response regression $\updownarrow$
- Group selection in univariate response regression.


No rank and row sparse models; no adaptive methods tailored to both.
Joint rank and row selection: JRRS

- Will develop new criteria, for joint rank and predictor selection.
- $r \leq n \land |J|$, $\text{rank}(X) = q \leq m \land p$; $|J| \leq p$; $r$ and $J$ unknown.
- Optimal risk rates achievable adaptively by the G-Lasso, RSC/NNP and (to show) JRRS.

<table>
<thead>
<tr>
<th>Method</th>
<th>Complexity</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-Lasso</td>
<td>$</td>
<td>J</td>
</tr>
<tr>
<td>RSC or NNP</td>
<td>$(p + n)r$</td>
<td>in rank-sparse models</td>
</tr>
<tr>
<td>JRRS</td>
<td>$(</td>
<td>J</td>
</tr>
</tbody>
</table>

- JRRS rates never worse and typically much better.
A penalized least squares estimator

- $Y$ is a $m \times n$ matrix; $X$ is a $m \times p$ matrix.
- $\|M\|_F^2$ is the sum of the squared entries of $M \in \mathcal{M}_{p \times n}$.
- Candidate model $B \in \mathcal{M}_{p \times n}$ has number of parameters
  $$(n + |J(B)| - \text{rank}(B))\text{rank}(B) \leq (n + |J(B)|)\text{rank}(B).$$

The one-step JRRS estimator

$$\hat{A} = \arg\min_{B \in \mathcal{M}_{p \times n}} \{\|Y - XB\|_F^2 + c\sigma^2(2n + |J(B)|)\text{rank}(B)\}.$$ 

- Generalizes to multivariate response models
  the AIC/$C_p$-type criteria developed for univariate response.
More on the one-step JRRS penalty

- \( B \in \mathcal{M}_{p \times n} \) with \( J(B) \) non-zero rows.

- **JRRS penalty**  \( \text{pen}(B) \propto \sigma^2 (n + |J(B)|) \text{rank}(B) \)

- \( B \in \mathcal{M}_{p \times n} \) (ignoring non-zero rows), \( \text{rank}(X) = q \).

- **RSC penalty**  \( \text{pen}(B) \propto \sigma^2 (n + q) \text{rank}(B) \)

- Squared "error level" in full model = \( \mathbb{E}d_1^2(PE) \approx \sigma^2 (n + q) \),
  \( E \) with iid sub-Gaussian entries, \( P = X(X'X)^{-1}X' \).

- JRRS generalizes RSC to allow for variable selection.

- To reduce rank *and* select variables work with:
  \( \mathbb{E}d_1^2(P_{J(B)}E) \approx \sigma^2 (n + |J(B)|) \).
Oracle-type bounds for the risk of the one-step JRRS

- \( \text{rank}(A) = r \), non-zero rows of \( A \) with indices in \( J(A) = J \).

Adaptation to Row and Rank Sparsity via one-step JRRS

For all \( A \) and \( X \)

\[
\mathbb{E} \left[ \| XA - X\widehat{A} \|_2^2 \right] \lesssim \inf_B \left[ \| XA - XB \|_2^2 + \sigma^2(n + |J(B)|)r(B) \right] \\
\lesssim \sigma^2 \{ n + |J| \} r.
\]

- RHS = the best bias-variance trade-off across \( B \).
- \( \widehat{A} \) is adaptive: it mimics the behavior of an optimal estimator computed knowing \( r \) and \( J \).
- Minimax rate, under suitable conditions.
- Bound valid for any \( m, n, p \).
Select the best from a finite list

- If $p > 20$, JRRS estimation over all $B$ becomes computationally intractable
- $\mathcal{B} = \{B_1, \ldots, B_L\} =$ Finite (large) collection of (random) matrices with different sparsity patterns; may depend on data $X$ and $Y$.

**Optimal selection from a finite list via JRRS**

For all $A$ and $X$

$$
\mathbb{E} \left[ \|XA - X\tilde{A}\|^2 \right] \lesssim \inf_{1 \leq j \leq L} \left[ \|XA - XB_j\|^2 + \sigma^2 (n + J(B_j)) r(B_j) \right].
$$

$$
\tilde{A} = \arg \min_{B \in \mathcal{B}} \{\|Y - XB\|^2_F + c\sigma^2 (2n + |J(B)|) \text{rank}(B)\}.
$$
Rank Constrained Group Lasso: main building block

- One-step JRRS penalty $\text{pen}(B) \propto (n + |J(B)|) \text{rank}(B)$. $J(B)$ forces complete enumeration; for large $p$ that’s a problem!
- Idea: use convex relation $\|B\|_{2,1} = \sum_{j=1}^{p} \|b_j\|_2$.
- Set $\lambda_k \propto \sigma \sqrt{kd_1^2(X)}$, for each $k$.

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

- $\hat{B}_k$ is a Rank-Constrained G-Lasso. (RCGL) Other ”group” penalties possible.
\[ \hat{B}_k = \arg \min_{\text{rank}(B) \leq k} \left\{ \| Y - XB \|_F^2 + \lambda_k \| B \|_{2,1} \right\}. \]

- For \( k = n \wedge p \), estimator \( \hat{B}_k \) is G-Lasso.
- For \( \lambda = 0 \), estimator \( \hat{B}_k \) is a reduced-rank estimator.
- Otherwise, \( \hat{B}_k \) is a synthesis of the two; new algorithm needed. Efficient algorithm Bunea, She and Wegkamp (2011).
- Works in high dimensions.
Two-step JRRS: Method 1

**Method 1**

- **Step 1.** Use the Rank Selection Criterion RSC to estimate consistently $r$ by $\hat{r}$.

- **Step 2.** Compute the Rank Constrained G-Lasso estimator $\hat{B}_k$ with $k = \hat{r}$ to obtain the final estimator $\hat{B} = \hat{B}_{\hat{r}}$.

**Major Practical Advantage:** Easy tuning, backed up by theory.

- For Step 1: Same tuning parameter of RSC gives best $MSE$ and correct rank. Can use CV safely; other alternatives exist.
- For Step 2: We want best $MSE$, CV safe.
Two-step JRRS: Method 2

**Method 2**

- **Step 1.** Pre-specify a grid of values $\Lambda$ for $\lambda$. Use RCGL to construct
  
  $$
  B = \{\hat{B}_{k,\lambda} : k \in \{1, \ldots, q\}, \lambda \in \Lambda\}.
  $$

- **Step 2.** Compute
  
  $$
  \tilde{B} = \arg\min_{B \in B} \{\| Y - XB \|_F^2 + \text{pen}(B) \},
  $$

  with $\text{pen}(B) \propto \sigma^2(n + |J(B)| \text{rank}(B))$.

- Requires a 2-D grid search: more computationally involved than Met. 1.
Framework and motivation
Joint Rank and Row Selection Methods
Two-step JRRS estimators
Numerical performance and examples
Summary

Oracle-type bounds for the risk of the two-step JRRS

- Method 1 (RSC + RCGL) $\rightarrow \hat{B}$;  Method 2 (RCGL + AIC-M) $\rightarrow \tilde{B}$

Adaptation to Row and Rank Sparsity via two-step JRRS

For all $A$ and for $X$ satisfying Assumption 1

$$
\mathbb{E} \left[ \|XA - X\tilde{B}\|^2 \right] \lesssim \inf_B \left[ \|XA - XB\|^2 + \sigma^2 (n + J(B))r(B) \right] \\
\lesssim \sigma^2 \{n + J(A)\} r(A).
$$

If, in addition, $d_r(XA) > 2\sqrt{2}\sigma(\sqrt{n} + \sqrt{q})$, same inequality holds for $\hat{B}$.

- RHS = the best bias-variance trade-off across all matrices $B$.
- $\hat{B}$, $\tilde{B}$ are adaptive: mimic the behavior of an optimal estimator computed knowing $r(A)$ and $J(A)$.
- Bound valid for any $m, n, p$; computationally efficient.
Mild conditions on the design matrix

Assumption 1

There exists a set $J \subset \{1, \ldots, p\}$ and a number $\delta_J > 0$ such that

$$\frac{1}{m} \|XB\|^2_F \geq \delta_J \sum_{j \in J} \|b_j\|^2_2, \quad \text{for all } B = [b_1 \cdots b_p]^T \in \mathbb{R}^{p \times n}$$

- Only a sub-matrix of $X'X$ has a non-zero smallest eigen-value. Mild condition.
Large \( p \) - small \( m \) numerical performance comparison

- \( m = 30, |J| = 15, p = 100, n = 10, r = 2, \sigma^2 = 1. \)
- Performance comparison between:
  - rank and row reduction via \( \text{RSC} \rightarrow \text{RCGL} \) and \( \text{G-LASSO} \rightarrow \text{RSC} \),
  - only row via \( \text{G-LASSO} \), and only rank via \( \text{RSC} \).
- All optimally tuned on a very large independent set.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{RSC} \rightarrow \text{RCGL} )</td>
<td>363</td>
</tr>
<tr>
<td>( \text{G-LASSO} \rightarrow \text{RSC} )</td>
<td>402</td>
</tr>
<tr>
<td>( \text{G-LASSO} )</td>
<td>511</td>
</tr>
<tr>
<td>( \text{RSC} )</td>
<td>1905</td>
</tr>
</tbody>
</table>
Large $m$ - small $p$ numerical performance comparison

- $m = 100$, $|J| = 15$, $p = 25$, $n = 25$, $r = 5$, $\sigma^2 = 1$.
- Performance comparison between:
  rank and row reduction via $\text{RSC} \rightarrow \text{RCGL}$, $\text{G-LASSO} \rightarrow \text{RSC}$, only row via $\text{G-LASSO}$, and only rank via $\text{RSC}$
- All optimally tuned on a very large independent set.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{RSC} \rightarrow \text{RCGL}$</td>
<td>8.1</td>
</tr>
<tr>
<td>$\text{G-LASSO} \rightarrow \text{RSC}$</td>
<td>8.1</td>
</tr>
<tr>
<td>$\text{RSC}$</td>
<td>11.5</td>
</tr>
<tr>
<td>$\text{G-LASSO}$</td>
<td>17.7</td>
</tr>
</tbody>
</table>
A study of the effect of HIV-infection on human cognitive abilities

- HIV-Neuroimaging laboratory at Brown University, PI R. Cohen.
- \( m = 62 \) HIV+ patients, also infected with Hepatitis C, and with a history of drug abuse
- \( n = 13 \) neuro-cognitive indices (NCIs) from five domains: attention/working memory, speed of information processing psychomotor abilities, executive function, and learning and memory.
- \( p = 234 \) predictors (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.
RSC and JRRS: two rank-1 models

- Both methods: One new predictive score $S$.
- Left = RSC; $MSE = 193$; $S = \text{lin. comb. of } p = 234$ predictors.
- Right = JRRS; $MSE = 138$; $S = \text{lin. comb. of } |J| = 10$ predictors.
• JRRS selected rank 1 and only 10 predictors.
• Education is one of them, confirming past findings.
• The fractional anisotropy at corpus callosum stands out among the very many DTI-derived measures, in terms of predictive power.
• New finding in the lab and first quantitative confirmation.
## Summary

<table>
<thead>
<tr>
<th>Methods</th>
<th>Adaptation to RR-sparsity</th>
<th>Assumptions on X and/or A</th>
<th>Restrictions on p</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-step JRRS (AIC-M)</td>
<td>Yes</td>
<td>None</td>
<td>$p \leq 20$</td>
</tr>
<tr>
<td>Two-step JRRS1 (RSC → RCGL)</td>
<td>Yes</td>
<td>Restricted Eigenvalue; $d_r(XA) \gg \text{&quot;noise level&quot;}$</td>
<td>None</td>
</tr>
<tr>
<td>Two-step JRRS2 (RCGL → AIC-M)</td>
<td>Yes</td>
<td>Restricted Eigenvalue</td>
<td>None</td>
</tr>
<tr>
<td>GL → RSC</td>
<td>Yes</td>
<td>Mutual coherence et al. $\min_j |a_j|_2 \gg \text{noise level}$</td>
<td>None</td>
</tr>
</tbody>
</table>

- **RSC → RCGL** easy to tune in practice; backed up by theory. Best!
- **RCGL → AIC-M** tuning requires search over a 2-D grid. Second best!
- **GL → RSC**: (1) Most restrictive theoretical assumptions;
  (2) Requires tuning for consistent group selection, open problem!
Summary: Our contribution

Jointly rank and row-sparse models and their estimation

1. Introduced jointly rank and row sparse models.
2. Offered new procedures tailored to the new class of models.
3. Showed that the one-step JRRS is a theoretically optimal adaptive procedure:
   Finite sample oracle inequalities for $\mathbb{E}\|XA - X\hat{A}\|_F^2$ for all $A$ and $X$.
4. Introduced computationally efficient two-step JRRS.
5. Two-step JRRS satisfy finite sample oracle inequalities under minimal conditions on $X$.
6. Guaranteed small $\mathbb{E}\|XA - X\hat{A}\|_F^2$ if $A$ of low rank and few non-zero rows. Analysis valid for all $m$, $n$, $p$, rank $r$ and $|J|$. In particular, $r$ and $|J|$ can grow with $m$ and $n$. 
Talk based on

- Florentina Bunea, Yiyuan She and Marten Wegkamp

- Florentina Bunea, Yiyuan She and Marten Wegkamp

- Research partially supported by NSF-DMS 1007444.