Optimal selection of reduced rank estimators of high-dimensional matrices

Marten Wegkamp

Department of Statistics, Florida State University

14 January 2011, Gainesville

Talk based on:

Florentina Bunea, Yiyuan She and Marten Wegkamp.
Optimal selection of reduced rank estimators of high-dimensional matrices.
arXiv:1004.2995v1, 18 April 2010. To appear in AoS.

Multivariate Response Regression Model

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

$$Y = XA + E$$

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

Aim of Our Study

The aim is to estimate a low-rank approximation of A.

- 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.
- Estimators restricted to have rank equal to a fixed number $k \leq n \wedge p$ were introduced to remedy this drawback.

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.

There are no theoretical results on the properties of the selected reduced rank estimate.

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 $P = X(X'X)^{-}X'$.

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 both $\widehat{A}$ and $\widehat{k} = r(\widehat{A})$ have a closed form solution and can be efficiently computed based on the SVD of PY.

Proposition

- $\widehat{k}$ is the number of singular values of PY that exceed $\sqrt{\mu}$
- $\widehat{A}$ is the rank restricted LSE (of rank $\widehat{k}$)

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 $\widehat{B}_k = W_k G_k$.

Consistent Effective Rank Estimation

Theorem

Suppose that there exists an index $s \leq r$ 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

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

- We can consistently estimate the index s provided we use a large enough value for μ 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 μ, 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

$$\mu = C_0\sigma^2(\sqrt{q} + \sqrt{n})^2$$

as our measure of the noise level, for some $C_0 > 1$.

Summarizing,

Corollary

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

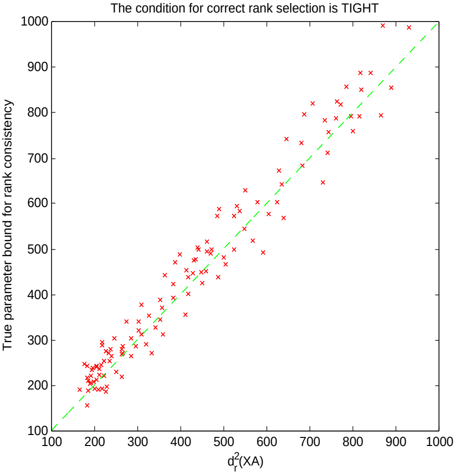

Figure: Tightness of the consistency condition

Risk Bounds for the Restricted Rank LSE

Theorem

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

$$\|X\widehat{B}_k - XA\|_F^2 \leq 3\left[\sum_{j>k} d_j^2(XA) + 4kd_1^2(PE)\right]$$

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 μ,

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

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.

Unknown σ^2

Theorem

For large $n+q$ and large $n(m-q)$ and

$$\text{pen}(B) = C_0(\sqrt{n}+\sqrt{q})^2 \frac{\|Y - PY\|_F^2}{mn - qn} r(B),$$

we have

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

Nuclear Norm Penalized Estimators

We compare our RSC estimator $\widehat{A}$ with the alternative estimator $\widetilde{A}$ that minimizes

$$\|Y - XB\|_F^2 + 2\tau\|B\|_1$$

over all $p \times n$ matrices B.

Theorem

On the event $d_1(X'E) \leq \tau$, we have, for any B,

$$\|X\widetilde{A} - XA\|_F^2 \leq \|XB - XA\|_F^2 + 4\tau\|B\|_1.$$

Nuclear Norm Penalized Estimators

Theorem

For $\tau = (1+\theta)\sigma d_1(X)(\sqrt{n}+\sqrt{q})$,

$$\mathbb{P}\left\{\|X\widetilde{A} - XA\|_F^2 \le \|XB - XA\|_F^2 + 4\tau\|B\|_1\right\}$$
$$\ge 1 - \exp\left\{-\frac{1}{2}\theta^2(n+q)\}\right\}$$

- It is possible to obtain an oracle inequality for $\widetilde{A}$ that resembles the oracle inequality for $\widehat{A}$.
- Our bounds for $\widehat{A}$ are much cleaner and obtained under fewer restrictions on the design matrix.
- We need that the condition number $c_0(X'X) = \lambda_1(X'X)/\lambda_p(X'X)$ is finite.
- Proof uses arguments similar to Negahban and Wainwright (2009) and Rohde and Tsybakov (2010)
- NNP fails to select the correct rank.

Rank Recovery

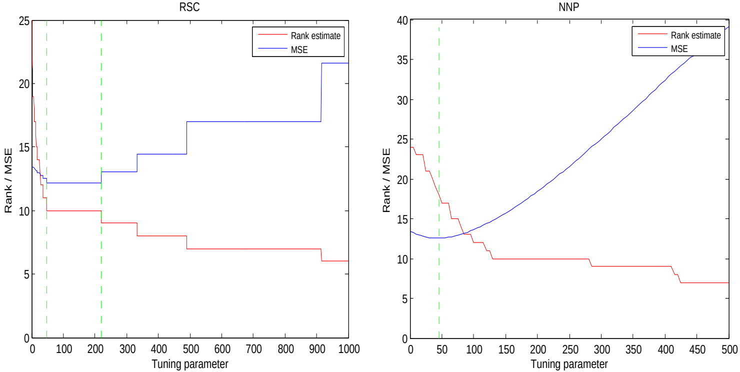

Figure: The MSE and rank of the estimators RSC (left) and NNP (right) as a function of the tuning parameter.

Rank Recovery

We suggest

$$\tilde{k} = \max\{k : \; d_k(X'X\widetilde{A}) > 2\tau\}.$$

Theorem

Let $r = r(A)$ and assume that $d_r(X'XA) > 4\tau$. Then

$$\begin{aligned} \mathbb{P}\{\tilde{k} \neq r\} &\leq \mathbb{P}\{d_1(X'E) > \tau\} \\ &\leq \exp\left\{-\frac{1}{2}\theta^2(n+q)\right\} \end{aligned}$$

for $\tau = (1+\theta)\sigma d_1(X)(\sqrt{n} + \sqrt{q})$.

Simulations

- RSC with $\mu = 2S^2(n+q)$.
- $X = [x_1, x_2, \cdots, x_m]'$ by generating its rows x_i i.i.d. from MVN$(\mathbf{0}, \Sigma)$, with $\Sigma_{jk} = \rho^{|j-k|}$, $\rho > 0$, $1 \le j, k \le p$.
- $A = bB_0B_1$, with $b > 0$, B_0 is a $p \times r$ matrix and B_1 is a $r \times n$ matrix. All entries in B_0 and B_1 are i.i.d. $N(0,1)$.
- Each row in $Y = [y_1, \cdots, y_m]'$ is then generated as $y_i = x_i'A + E_i$, $1 \le i \le m$, with E_i the i-th row of E with $N(0,1)$ i.i.d. entries.

- Each simulated model is characterized by the following control parameters : m (sample size), p (number of independent variables), n (number of response variables), r (rank of A), ρ (design correlation), and b (signal strength).
- Experiment 1: number of predictors $p <$ sample size m. $m = 100, p = 25, n = 25, r = 10$, correlation coefficient $\rho = 0.1, 0.5, 0.9$ and signal strength $b = 0.1, 0.2, 0.3, 0.4$.
- Experiment 2: $p > m$. $m = 20, p = 100, n = 25, r = 10$, correlation $\rho = 0.1, 0.5, 0.9$ and signal strength $b = 0.1, 0.2, 0.3$

Performance comparisons of **Experiment 1**

		$\text{RSC}\|_{adap}$	$\text{RSC}\|_{val}$	$\text{NNP}\|_{val}$	$\text{NNP}^{(c)}\|_{val}$
$b = 0.1$					
$\rho = 0.9$	MSE(XA), MSE(A)	16.6, 5.3	16.3, 5.2	11.5, 3.0	16.5, 5.3
	RE, RRP	6, 0%	6, 0%	12, 0%	6, 0%
$\rho = 0.5$	MSE(XA), MSE(A)	18.7, 1.4	18.1, 1.4	16.2, 1.1	18.1, 1.4
	RE, RRP	8, 0%	9, 40%	16.5, 0%	9, 35%
$\rho = 0.1$	MSE(XA), MSE(A)	19.3, 1.0	18.0, 0.9	16.9, 0.8	18.0, 0.9
	RE, RRP	9, 0%	10, 75%	17, 0%	10, 65%
$b = 0.2$					
$\rho = 0.9$	MSE(XA), MSE(A)	18.4, 7.0	17.9, 7.1	15.9, 5.4	17.9, 7.1
	RE, RRP	8, 0%	9, 20%	16, 0%	9, 15%
$\rho = 0.5$	MSE(XA), MSE(A)	16.7, 1.3	16.7, 1.3	18.9, 1.5	16.7, 1.3
	RE, RRP	10, 100%	10, 100%	19, 0%	10, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	16.5, 0.9	16.5, 0.9	19.2, 1.0	16.5, 0.9
	RE, RRP	10, 100%	10, 100%	18, 0%	10, 100%

Performance comparisons of **Experiment 1**

		$\text{RSC}\|_{adap}$	$\text{RSC}\|_{val}$	$\text{NNP}\|_{val}$	$\text{NNP}^{(c)}\|_{val}$
$b = 0.3$					
$\rho = 0.9$	MSE(XA), MSE(A)	17.4, 7.0	17.3, 6.9	17.7, 6.7	17.3, 7.0
	RE, RRP	10, 65%	10, 95%	18, 0%	10, 80%
$\rho = 0.5$	MSE(XA), MSE(A)	16.4, 1.3	16.4, 1.3	19.8, 1.6	16.4, 1.3
	RE, RRP	10, 100%	10, 100%	19, 0%	10, 100 %
$\rho = 0.1$	MSE(XA), MSE(A)	16.4, 0.9	16.4, 0.9	19.9, 1.1	16.4, 0.9
	RE, RRP	10, 100%	10, 100%	19, 0%	10, 100%
$b = 0.4$					
$\rho = 0.9$	MSE(XA), MSE(A)	16.8, 6.6	16.8, 6.7	18.7, 7.4	16.8, 6.8
	RE, RRP	10, 100%	10, 100%	18, 0%	10, 85%
$\rho = 0.5$	MSE(XA), MSE(A)	16.3, 1.3	16.3, 1.3	20.3, 1.7	16.3, 1.3
	RE, RRP	10, 100%	10, 100%	20, 0%	10, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	16.3, 0.9	16.3, 0.9	20.3, 1.1	16.3, 0.9
	RE, RRP	10, 100%	10, 100%	20, 0%	10, 100%

Performance comparisons of **Experiment 2**

		$\text{RSC}\|_{adap}$	$\text{RSC}\|_{val}$	$\text{NNP}\|_{val}$	$\text{NNP}^{(c)}\|_{val}$
		$b = 0.1$			
$\rho = 0.9$	MSE(XA), MSE(A)	29.4, 3.9	29.4, 3.9	36.4, 3.9	29.4, 3.9
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.5$	MSE(XA), MSE(A)	29.1, 3.9	29.1, 3.9	37.2, 3.9	29.1, 3.9
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	29.0, 3.9	29.0, 3.9	37.2, 4.0	29.0, 3.9
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
		$b = 0.2$			
$\rho = 0.9$	MSE(XA), MSE(A)	28.9, 15.7	28.9, 15.7	38.7, 15.7	28.9, 15.7
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.5$	MSE(XA), MSE(A)	28.6, 15.7	28.6, 15.7	39.0, 15.7	28.6, 15.7
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	28.7, 15.8	28.7, 15.8	38.7, 15.8	28.7, 15.8
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
		$b = 0.3$			
$\rho = 0.9$	MSE(XA), MSE(A)	28.8, 35.3	28.8, 35.3	39.2, 35.3	28.8, 35.3
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.5$	MSE(XA), MSE(A)	28.5, 35.4	28.5, 35.4	39.5, 35.4	28.5, 35.4
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100 %
$\rho = 0.1$	MSE(XA), MSE(A)	28.6, 35.5	28.6, 35.5	39.3, 35.5	28.6, 35.5
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%

Conlcusions of our Simulation Study

1. RSC with adaptive choice performs well - as well as with optimally tuned μ.
2. For moderate or high SNR$= d_r(XA)/(\sqrt{n}+\sqrt{q})$ and for low to moderate correlation between the predictors, RSC has excellent behavior.
3. For low SNR, or for high correlation between some covariates, NNP may be slightly more accurate than the RSC.
4. The correct rank, 10, is always overestimated by NNP.
5. A two-staged estimator, NNP$^{(c)}$, provides a successful improvement over NNP, for rank selection.
6. RSC is much more computationally efficient than NNP$^{(c)}$.

Summary: Our Contribution

- RSC criterion is easy to compute (closed form).
- Appropriate notion of signal and noise.
- Correct rank identification.
- Finite sample oracle inequalities for fit of $X\widehat{A}$ for all A and X.
- Finite sample analysis valid for all m, n, p and rank r.
- NNP has similar theoretical properties, under more stringent conditions on X. NNP is not the most parsimonious estimator.

Thanks!