

Estimation of large dimensional sparse covariance matrices

Noureddine El Karoui

Department of Statistics
UC, Berkeley

May 15, 2009

Sample covariance matrix and its eigenvalues

- Data: $n \times p$ matrix X
- n (independent identically distributed) observations of a random vector $(X_i)_{i=1}^n$ in \mathbb{R}^p .
- Suppose X_i -s have covariance matrix Σ_p
- Often interested in estimating Σ_p (e.g for PCA); eigenvalues: λ_i
- Standard estimator: $\hat{\Sigma}_p = (X - \bar{X})(X - \bar{X})' / (n - 1)$; eigenvalues l_i

- Principal Component Analysis (**PCA**); *large* eigenvalues:
- **Optimization problems:** Example in risk management:
“Invest” a_j in $X_{t,j}$. Risk measured by $\text{var}(X_t a) = a' \Sigma_p a$.
(Constraints on a in more realistic versions: e.g $\sum a_i = 1$)
Role of *small*(/all) eigenvalues.

Classical results from multivariate Statistics

Classical theory: p fixed, $n \rightarrow \infty$

Fact: $\hat{\Sigma}$ unbiased, \sqrt{n} -consistent estimator of Σ :

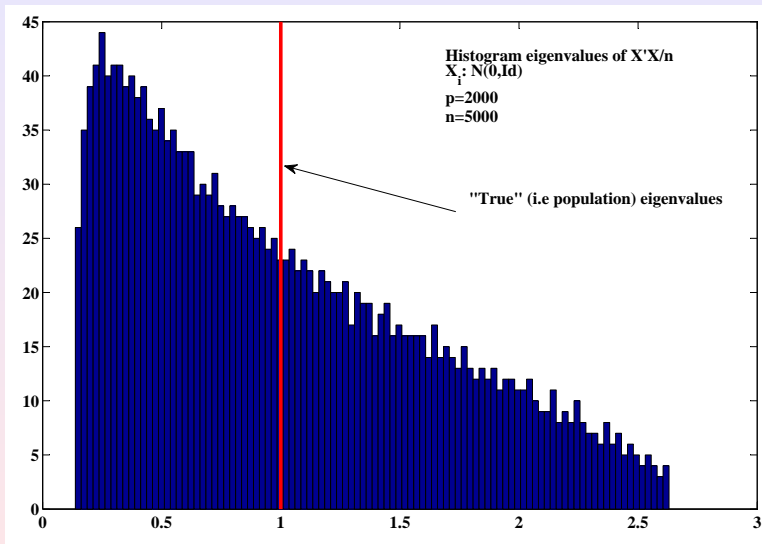
$$\sqrt{n}\|\hat{\Sigma} - \Sigma\| = O_P(1) .$$

Also, fluctuation theory (CLT) for largest eigenvalues (Anderson, '63), under eg normality of the data

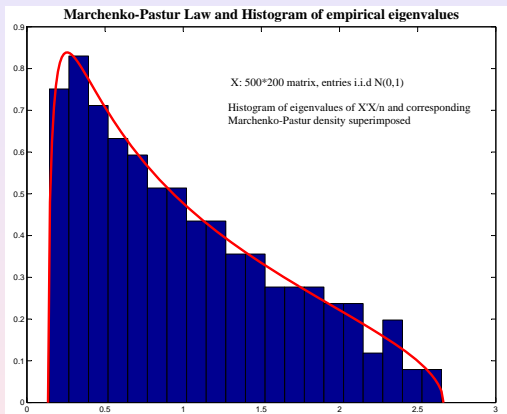
Much more is known about estimation of covariance matrices: Efron, Haff, Morris, Stein etc...

Will focus here on “large n , large p ” situation (quite common now), so p/n has finite non-zero limit

Visual Example



Graphical illustration: $n=500, p=200$



Note: $\Sigma = \text{Id}$, so all population (or “true”) eigenvalues equal 1.

Surprising?

- Previous plots perhaps surprising:
- In particular, CLT implies that $\hat{\sigma}_{ij} - \sigma_{ij} = O(n^{-1/2})$: good entrywise estimation
- But poor spectral estimation

Importance of $\rho_n = p/n$

Case of $\Sigma = \text{Id}$

Suppose entries of $n \times p$ matrix X are iid mean 0, sd 1, 4th moment. **Population covariance = Id**

Let $\rho_n = p/n$.

Theorem (Marčenko-Pastur, '67)

l_i : (decreasing) eigenvalues of $\frac{1}{n}X'X$. Assume $\rho_n \rightarrow \rho \in (0, 1]$.
If $F_p(x) = \frac{1}{p}\{\#\{l_i \leq x\}\}$, then

$$F_p \Longrightarrow F_\rho \text{ in probability.}$$

F_ρ has known density.

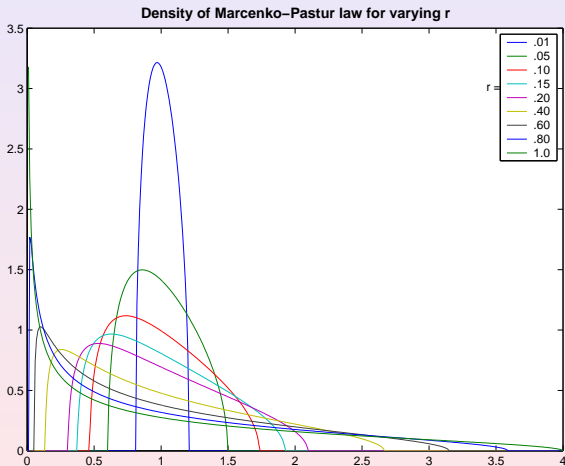
Support: $[a, b]$, with $a = (1 - \rho^{1/2})^2$, $b = (1 + \rho^{1/2})^2$

Bias issue/inconsistency problems for these asymptotics.

Marcenko-Pastur law illustration

Population covariance: $\Sigma = \text{Id}$

Here, density of limiting spectral distribution computable:



Remark about extensions/limitations

- RMT also gives results about limiting spectral distribution of eigenvalues when $\Sigma \neq \text{Id}$
- Models of form $X_i = \Sigma^{1/2} Y_i$, entries of Y_i i.i.d (say $4+\epsilon$ moments)
- Problem: geometric implications for the data; near orthogonality, near constant norm of X/\sqrt{p} , when $\lambda_1(\Sigma)$ bounded
- Two models can have same Σ and different limiting spectral distributions

Previous results highlights fact that naive estimation of covariance matrix results in inconsistency in high-dimension.

Question: can we find procedures that consistently estimate spectral properties of these matrices? (I.e also eigenspaces) And are more robust to distributional assumptions?

Can we exploit good entrywise estimation to improved spectral estimation?

Previous work

Related ideas: Banding

Scheme proposed by P. Bickel and L. Levina ('06):

- Consider population covariance matrices with small entries away from the diagonal
- Technically: Σ_p satisfies

$$\max_j \left\{ \sum_{|i-j|>m} |\sigma(i,j)| \right\} < Cm^{-\alpha}, \forall m$$

- Recommend **banding**: $\hat{\sigma}_{i,j} = 0$ if $|i - j| > k$, otherwise keep estimate from $\hat{\Sigma}_p$
- Call this estimator $B_k(\hat{\Sigma}_p)$

Theorem (Bickel and Levina)

If $k \asymp (n^{-1/2} \log(p))^{-1/(\alpha+1)}$, + tail decay conditions

$$\|B_k(\hat{\Sigma}_p) - \Sigma_p\|_2 \rightarrow 0$$

- Oft-made assumption: many $\sigma(i, j) = 0$ or “small”
- Appeal of thresholding (i.e keep large values, and put small ones to 0)
- Aim: find simple methods for improving estimation
- Practical requirements: speed, parallelizability, limited assumptions
- Theoretical requirements: good convergence properties, in spectral norm ($\| \cdot \|_2$)
- Estimator not sensitive to ordering of variables

What is a good estimator?

Requirement:

$$\| \hat{\Sigma}_p - \Sigma_p \|_2 \rightarrow 0$$

where (for symmetric matrices)

$$\| M \|_2 = \sigma_1(M) = \sqrt{\lambda_1(M^*M)} = \max_i |\lambda_i(M)| .$$

Convergence of $\| \cdot \|_2$ implies

- Consistency of **all** eigenvalues (Weyl's inequality)
- Consistency of stable subspaces corresponding to **separated** eigenvalues (Davis-Kahan $\sin \theta$ theorem)

Our aim: permutation equivariant estimator, operator-norm consistent

Our strategy: (hard) thresholding

Standard notion of sparsity

Ill-suited for spectral problems

Standard notion : count number of non-zero elements

$$E_1 = \begin{pmatrix} 1 & \frac{1}{\sqrt{p}} & \frac{1}{\sqrt{p}} & \cdots & \frac{1}{\sqrt{p}} \\ \frac{1}{\sqrt{p}} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \frac{1}{\sqrt{p}} & 0 & 0 & 1 & 0 \\ \frac{1}{\sqrt{p}} & 0 & 0 & \cdots & 1 \end{pmatrix} \quad E_2 = \begin{pmatrix} 1 & \frac{1}{\sqrt{p}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{p}} & 1 & \frac{1}{\sqrt{p}} & \cdots & 0 \\ 0 & \frac{1}{\sqrt{p}} & 1 & \frac{1}{\sqrt{p}} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{\sqrt{p}} \\ 0 & \cdots & 0 & \frac{1}{\sqrt{p}} & 1 \end{pmatrix}$$

Eigenvalues E_1 :

$$1 \pm \sqrt{\frac{p-1}{p}}, 1$$

(multiplicity: $p-2$)

Eigenvalues E_2 :

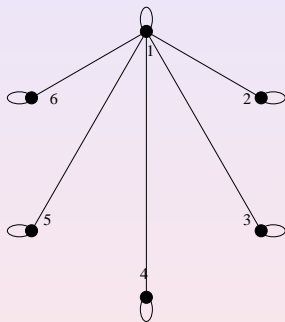
$$1 + \frac{2}{\sqrt{p}} \cos\left(\frac{\pi k}{p+1}\right),$$

$$k = 1, \dots, p$$

Adjacency matrices

$$E_1 = \begin{pmatrix} 1 & \frac{1}{\sqrt{p}} & \frac{1}{\sqrt{p}} & \cdots & \frac{1}{\sqrt{p}} \\ \frac{1}{\sqrt{p}} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \frac{1}{\sqrt{p}} & 0 & 0 & 1 & 0 \\ \frac{1}{\sqrt{p}} & 0 & 0 & \cdots & 1 \end{pmatrix}$$

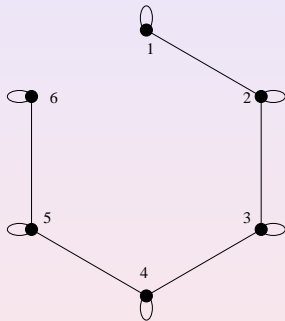
$$A_1 = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$



Adjacency matrices

$$E_2 = \begin{pmatrix} 1 & \frac{1}{\sqrt{p}} & 0 & \dots & 0 \\ \frac{1}{\sqrt{p}} & 1 & \frac{1}{\sqrt{p}} & \dots & 0 \\ 0 & \frac{1}{\sqrt{p}} & 1 & \frac{1}{\sqrt{p}} & \dots \\ \vdots & \ddots & \ddots & \ddots & \frac{1}{\sqrt{p}} \\ 0 & \dots & 0 & \frac{1}{\sqrt{p}} & 1 \end{pmatrix}$$

$$A_2 = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ 0 & 1 & 1 & 1 & \dots \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & \dots & 0 & 1 & 1 \end{pmatrix}$$

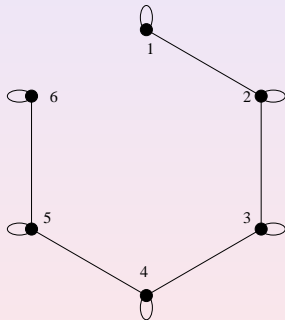
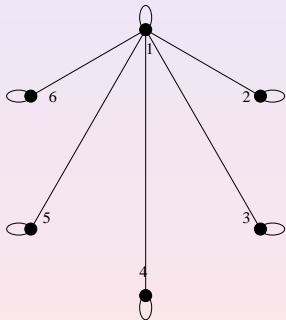


Adjacency matrices and graphs: comparison

Closed walks of length k

Closed walk: start and finishes at same vertex

Length of walk: number of vertices it traverses

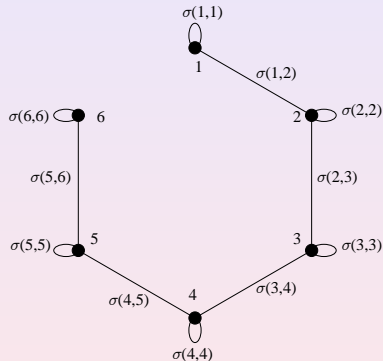
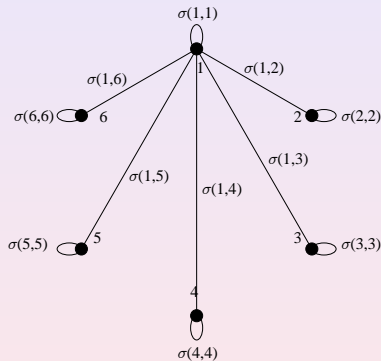


Adjacency matrices and graphs

Connection with spectrum of covariance matrices

Closed walk (length k) $\gamma: i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_k \rightarrow i_{k+1} = i_1$

Weight of $\gamma: w_\gamma = \sigma(i_1, i_2) \dots \sigma(i_k, i_1)$



$$\text{trace}(\Sigma^k) = \sum \lambda_i^k(\Sigma) = \sum_{\gamma \in \mathcal{C}_p(k)} w_\gamma$$

Notion of sparsity compatible with spectral analysis

A proposal

- Given Σ_p , $p \times p$ covariance matrix, compute adjacency matrix $A_p = 1_{\sigma(i,j) \neq 0}$
- Associate graph \mathcal{G}_p to it
- Consider $\mathcal{C}_p(k) =$
 {closed walks of length k on the graph with adjacency matrix A_p }
 and $\phi_p(k) = \text{Card} \{ \mathcal{C}_p(k) \} = \text{trace} (A_p^k)$.

Call sequence of Σ_p **β -sparse** if

$$\forall k \in 2\mathbb{N}, \phi_p(k) \leq f(k)p^{\beta(k-1)+1}$$

where $f(k)$ independent of p and $0 \leq \beta < 1$

Examples

Computation of sparsity coefficients

- **Diagonal matrix** : $A_p = \text{Id}_p$. $\phi(k) = p$, for all k . Sparsity coefficient: $\mathbf{0}$.
- **Matrices with at most M non-zero elements on each line** $\phi(k) \leq pM^{k-1}$. Sparsity coefficient: $\mathbf{0}$.
- **Matrices with at most Mp^α non-zero elements on each line** $\phi(k) \leq M^{(k-1)}pp^{\alpha(k-1)}$. Sparsity coefficient: α

In all that follows,

- 1 $\Sigma_p(i, i)$ stay bounded
- 2 $X_{i,j}$ have infinitely many moments
- 3 Rows of $(n \times p)$ data matrix X i.i.d
- 4 $p/n \rightarrow l \in (0, \infty)$

Simple case: gap in entries of covariance matrix

Gaussian MLE, centered case

- $X_{i,j}$ centered; $\text{cov}(X_i) = \Sigma$
- Suppose Σ_p β -sparse, $\beta = 1/2 - \eta$ and $\eta > 0$

Theorem

$$S_p = \frac{1}{n} \sum_{i=1}^n X_i' X_i, \text{ and } T_\alpha(S_p)(i,j) = S_p(i,j) \mathbf{1}_{|S_p(i,j)| > Cn^{-\alpha}}.$$

$T_\alpha(S_p)$ = thresholded version of S_p at level $Cn^{-\alpha}$

Then, if $\alpha = \beta + \epsilon$, $\epsilon < \eta/2$,

$$\| \| T_\alpha(S_p) - \Sigma_p \| \|_2 \rightarrow 0 \text{ i.p.}$$

Beyond truly sparse matrices

Approximation by sparse matrices

How does thresholding perform on matrices approximated by sparse matrices?

- Suppose $\exists T_{\alpha_1}(\Sigma_p) = \tilde{\Sigma}_p$, β -sparse.
- Suppose $\|\tilde{\Sigma}_p - \Sigma_p\|_2 \rightarrow 0$.
- Suppose $\exists \alpha_0 < \alpha_1 < 1/2 - \delta_0$ such that adjacency matrix of (i, j) 's such that $Cn^{-\alpha_1} < |\sigma(i, j)| < Cn^{-\alpha_0}$ is γ -sparse, $\gamma \leq \alpha_0 - \zeta_0$, $\zeta_0 > 0$.

Proposition

Then conclusions of theorem above apply: for $\alpha \in (\alpha_0, \alpha_1)$,

$$\|T_\alpha(S_p) - \Sigma_p\|_2 \rightarrow 0 \text{ i.p.}$$

- Theorems apply to sample covariance matrix, i.e $(X - \bar{X})'(X - \bar{X})/(n - 1)$, not just the Gaussian MLE
- Also to correlation matrices
- Finite number of moments possible
- $p = n^r$ for certain r depending on β
- Proof provides finite n, p bounds on deviation probabilities of $\|T_\alpha(S_p) - \Sigma_p\|_2$

Example: $\Sigma_p^{(1)}(i, j) = \rho^{|i-j|}$, and $\Sigma_p = P' \Sigma_p^{(1)} P$, P random permutation matrix

- Can be approximated by $\tilde{\Sigma}_p = T_{n^{-1/2+\epsilon}}(\Sigma_p)$
- $\tilde{\Sigma}_p$ is asymptotically 0-sparse
- Proposition above applies when moment conditions satisfied

Sharpness of 1/2-sparse assumption

Consider

$$\Sigma_p = \begin{pmatrix} 1 & \alpha_2 & \alpha_3 & \dots & \alpha_p \\ \alpha_2 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \alpha_{p-1} & 0 & 0 & 1 & 0 \\ \alpha_p & 0 & 0 & \dots & 1 \end{pmatrix} .$$

with, e.g, $\alpha_i = \frac{1}{\sqrt{p}}$

Σ_p : 1/2-sparse (eigenvalues $A_p : (1 \pm \sqrt{p-1})$ and 1's)

Oracle estimator ($O(\hat{\Sigma}_p)$) of Σ_p inconsistent in Gaussian case:

$$\|O(\hat{\Sigma}_p) - \Sigma_p\|_2^2 = \sum_{i=2}^p (\hat{\alpha}_i - \alpha_i)^2$$

1/2 - η sparsity assumption "sharp"

Practical observations

Asymptopia?

Practical implementation

- Practicalities: implemented thresholding technique using FDR at level $1/\sqrt{p}$
- Theory: great results possible; Practice: good results (n, p a few 100's to a few 1000's); maybe not as good as hoped for.
- Practice very good when few coefficients to estimate; far away from " σ/\sqrt{n} " (unsurprisingly)
- Making "mistakes" OK. Softer techniques (Huber-type) might be good alternatives, especially for non-sparse matrices
- Eigenvector practice somewhat "disappointing". In hard situations, performs OK. In less hard situations, results comparable or a bit worse than sparse PCA.

Example thresholding

- Highlighted some difficulties in covariance estimation in “large n , large p ” setting
- Proposed notion of sparsity compatible with spectral analysis

Real world example

Data Analysis: Portfolio of Industry Indexes

Data:

- Daily Returns 48 Indexes, by Industry: Agriculture, Toys, Beer, Soda, etc... from Kenneth French's website at Dartmouth. $\mathbf{p} = 48$
- 2 years of observations $\mathbf{n} = 504$

Effect of shrinkage on smallest eigenvalue:

$$l_1^{\text{empirical}} \simeq 0.023$$

$$l_1^{\text{shrunk}} \simeq 0.217$$

Through Subsampling, get (0.205,0.303) 95% CI

Data Analysis: Portfolio of Industry Indexes

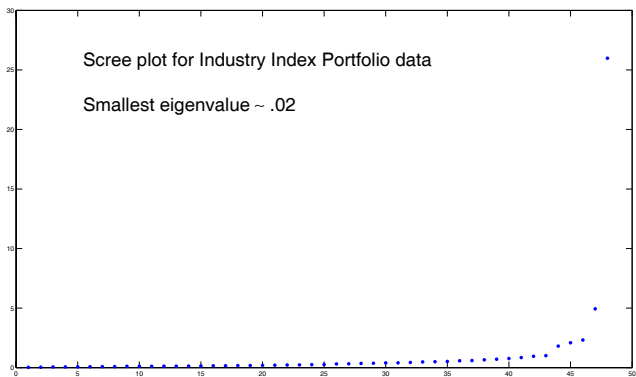


Figure: Scree plot for Industry index data

Data Analysis: Portfolio of Industry Indexes

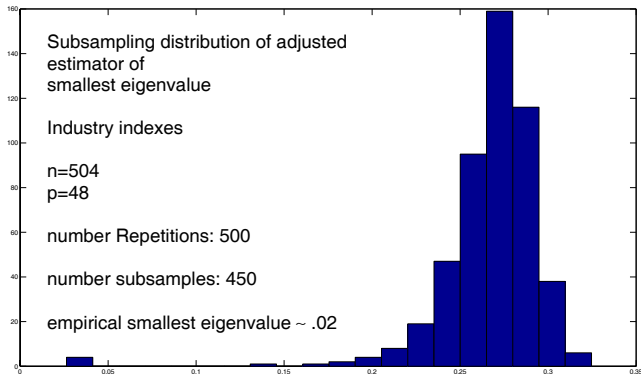
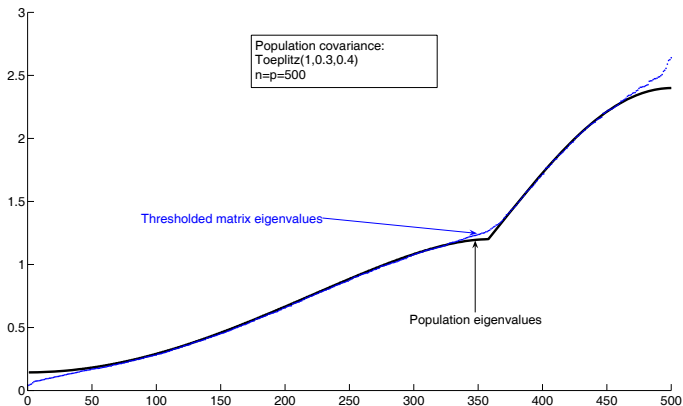


Figure: Subsampling distribution of estimator

Back

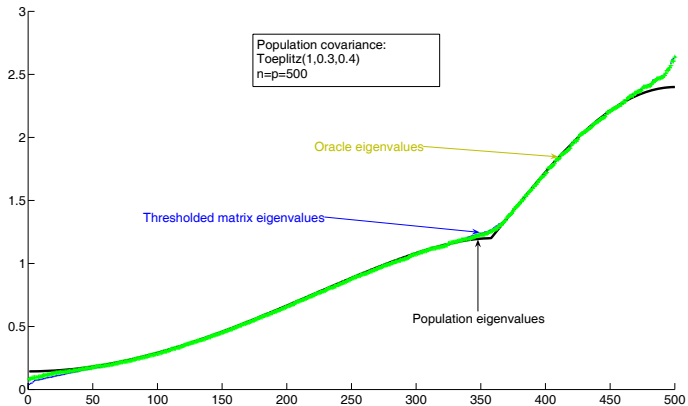
(Easy) Example 1: Toeplitz(1,.3,.4)

$n = p = 500$



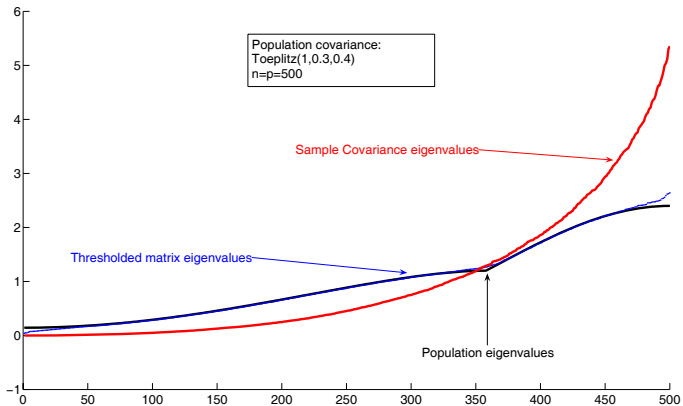
(Easy) Example 1: Toeplitz(1,.3,.4)

$n = p = 500$



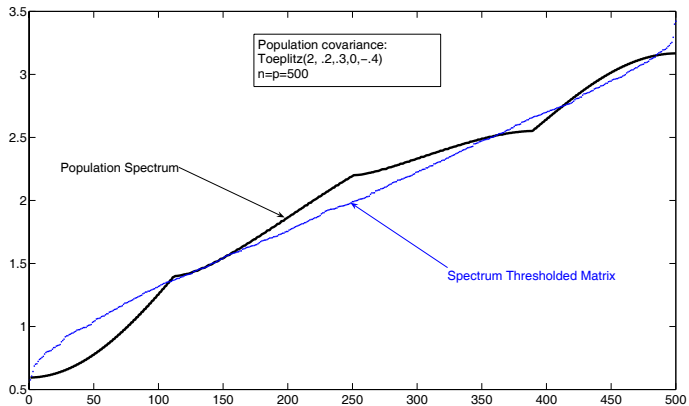
(Easy) Example 1: Toeplitz(1,.3,.4)

$n = p = 500$



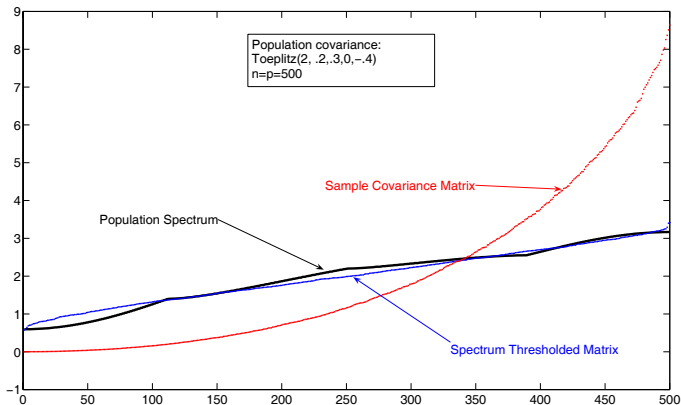
Example 2: Toeplitz(2,.2,.3,0,-.4)

$n = p = 500$



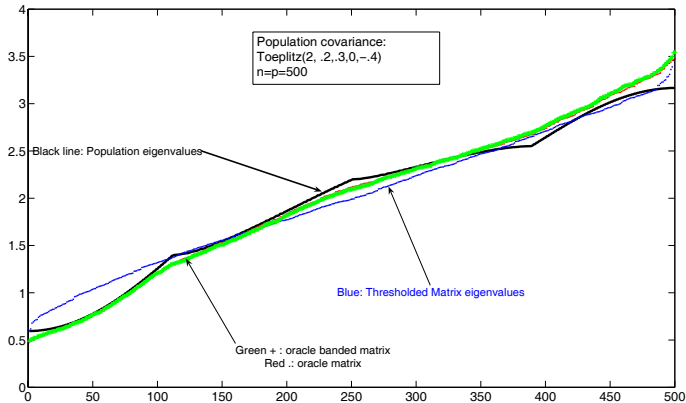
Example 2: Toeplitz(2,.2,.3,0,-.4)

$n = p = 500$



Example 2: Toeplitz(2,.2,.3,0,-.4)

Independent simulation



Back

Estimating the theoretical frontier

Practical implementation

- Data: from Fama-French website.
- Year 2005-2006. $n = 252$.
- Those are $p = 48$ industry indexes.

Orders of magnitude:

$$\alpha = -0.0401, \beta = 0.3449, \gamma = 14.2261 .$$

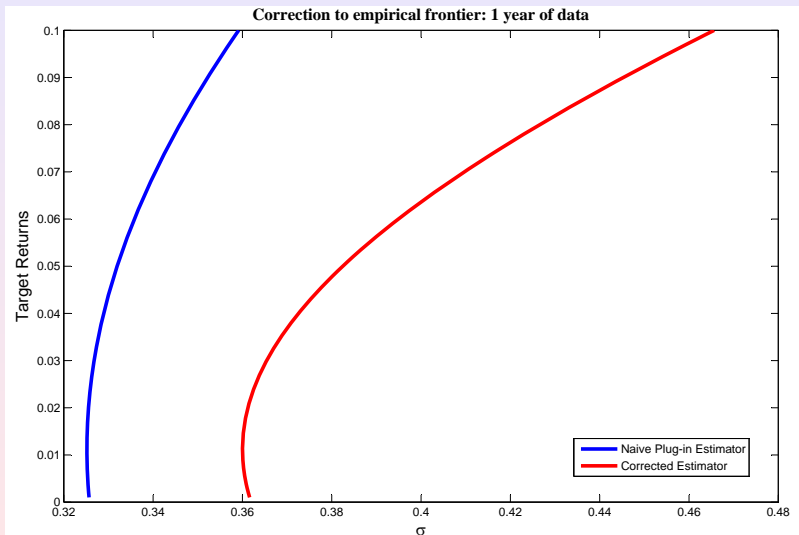
$$\hat{a} = -0.0324, \hat{b} = 0.0887, \hat{c} = 11.5163$$

Also,

$$\delta = 4.9050, \hat{d} = 1.0208$$

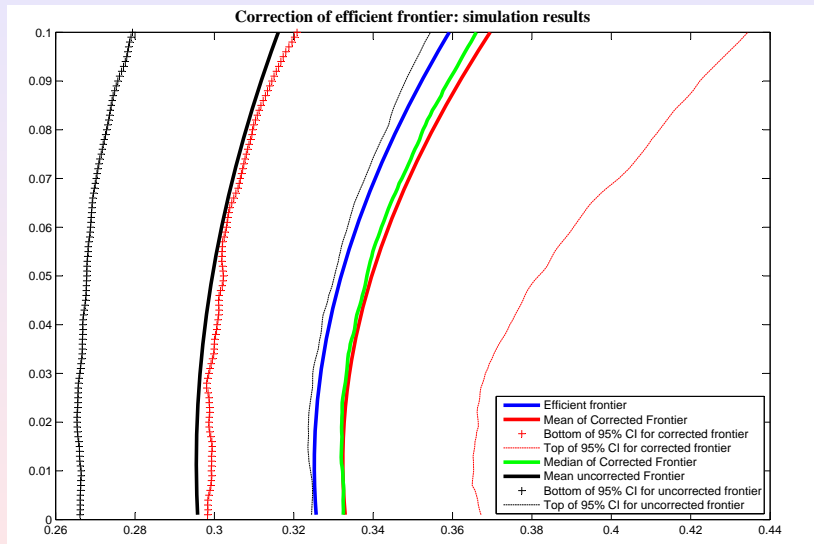
Estimating the theoretical frontier

Practical implementation: 252 days, 47 assets; raw data



Estimating the theoretical frontier

Practical implementation: 252 days, 47 assets; simulations



Eigenfaces

An interesting example of PCA

Question: compression of digitized pictures of human faces.

- Idea: Gather a database of faces.
- Example: ORL face database: 10 pictures of 40 distinct subjects
- Pictures are say $92 \times 112 = 10304$ pixels, 256 levels of gray images. Code them as vectors. Get data matrix X .
- Run Principal Component Analysis on the corresponding matrix
- Eigenvectors corresponding to large eigenvalues are “eigenfaces”.
- Need only a few numbers to approximate a new face: coefficients of its projection on eigenfaces.

Some eigenfaces for ORL face database

Data obtained from Wikipedia. Copyright by AT&T Laboratories Cambridge.

