Statistical inference in a spiked population model

Jian-feng YAO

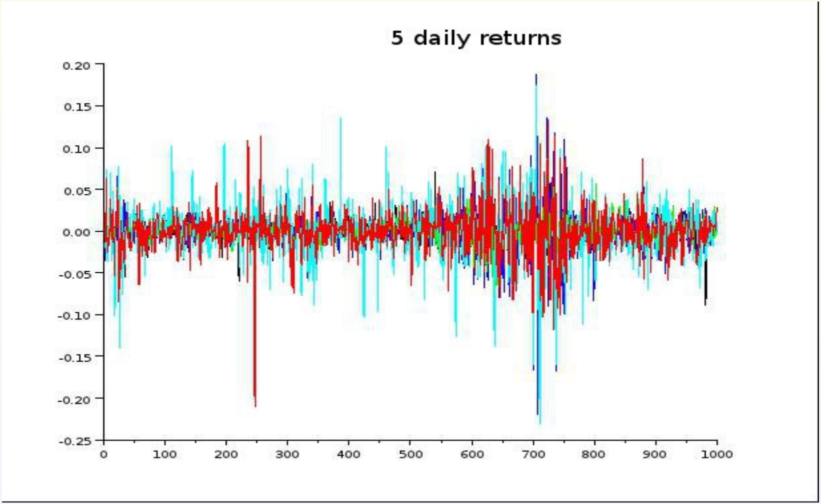
香港大學
THE UNIVERSITY OF HONG KONG

Joint work with Weiming LI (Beijing), Damien PASSEMIER (Rennes)

1. Spiked eigenvalues: an example
2. Inference on spikes: determination of their number q_0
   - Known results on spiked population
   - Estimator of q_0
   - Discussions on the estimator $\widehat{q}_0$
   - Application to S&P stocks data
3. Inference of the bulk spectrum
   - The problem and existing methods
   - A generalized expectation based method
   - Asymptotic properties of the GEE estimator
   - Application to S&P 500 stocks data

1) Spiked eigenvalues: an example

- SP 500 daily stock prices ; $p = 488$ stocks;
- $n = 1000$ daily returns $\mathbf{r}_t(i) = \log p_t(i)/p_{t-1}(i)$ from 2007-09-24 to 2011-09-12;



The sample correlation matrix

- Let the SCM (488×488)

$$S_n = \frac{1}{n}\sum_{t=1}^{n}(\mathbf{r}_t - \bar{\mathbf{r}})(\mathbf{r}_t - \bar{\mathbf{r}})^T \ .$$

- We consider the sample correlation matrix $\mathbf{R}_n$ with

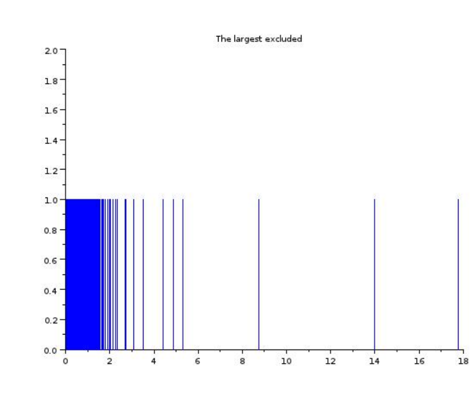
$$\mathbf{R}_n(i,j) = \frac{S_n(i,j)}{[S_n(i,i)S_n(j,j)]^{1/2}}.$$

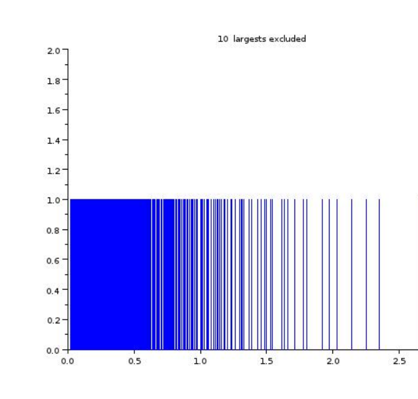
- The 10 largest and 10 smallest eigenvalues of $\mathbf{R}_n$ are:

237.95801	4.8568703	...	0.0212137	0.0178129
17.762811	4.394394	...	0.0205001	0.0173591
14.002838	3.4999069	...	0.0198287	0.0164425
8.7633113	3.0880089	...	0.0194216	0.0154849
5.2995321	2.7146658	...	0.0190959	0.0147696

Plots of sample eigenvalues

Left: 488 - 1 = 487 eigenvalues   right: 488 - 10 = 478 eigenvalues





$\Longrightarrow$ **the point:** **sample eigenvalues = bulk + spikes**

$\Longrightarrow$ **Analysis and estimation of spikes + bulk**

Random factor model

$$x_t = \sum_{k=1}^{q_0} a_k s_t(t) + \varepsilon_t = A s_t + \varepsilon_t,$$

- $s_t = (s_t(1), \ldots, s_t(q_0)) \in \mathbb{R}^{q_0}$ are $q_0 < p$ standardised random signals/factors,
- $A = (a_1, \ldots, a_{q_0})$, $p \times q_0$ deterministic matrix of factor loadings
- ε_t is an independent p-dimensional noise sequence, with a diagonal covariance matrix: $\Psi = \mathrm{cov}(\varepsilon_t) = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_p^2\}$.

Therefore,

$$\Sigma = \mathrm{cov}(x_t) = AA^* + \Psi \ .$$

- this model is very old; has wide range of application fields: psychology, chemometrics, signal processing, economics, etc.

2). Inference on spikes

a). Known results

Spiked population model

Population covariance matrix:

$$\Sigma = \mathrm{Cov}[\mathsf{x}_t] = \mathsf{AA}^* + \sigma^2 \mathsf{I}_p ,$$

with eigenvalues

$$\mathrm{spec}(\Sigma) = (\sigma^2 + \alpha'_1, \ \ldots, \ \sigma^2 + \alpha'_{q_0}, \underbrace{\sigma^2, \ \ldots, \ \sigma^2}_{p-q_0}) ,$$

where

- $\alpha'_1 \geq \alpha'_2 \geq \cdots \geq \alpha'_{q_0} > 0$ are non null eigenvalues of AA^*,

or equivalently

$$\mathrm{spec}(\Sigma) = \sigma^2 \times (\alpha_1, \ \ldots, \ \alpha_{q_0}, \underbrace{1, \ \ldots, \ 1}_{p-q_0}) ,$$

with

$$\alpha_i = 1 + \alpha'_i / \sigma^2 .$$

Asymptotic framework and assumptions

1. $p, n \to +\infty$ such that $p/n \to c$;
2. The population covariance matrix has K spikes $\alpha_1 > \cdots > \alpha_K$ with respective multiplicity numbers n_i, i.e.

$$\mathrm{spec}(\Sigma) = \sigma^2(\underbrace{\alpha_1, \ldots, \alpha_1}_{n_1}, \underbrace{\alpha_2, \ldots, \alpha_2}_{n_2}, \ldots, \underbrace{\alpha_K, \cdots, \alpha_K}_{n_K}, \underbrace{1, \cdots, 1}_{p-q_0});$$

$$[\ n_1 + \cdots + n_K = q_0\ ];$$

3. $\alpha_K > 1 + \sqrt{c}$ (detection level).
4. $\mathbb{E}(|x_{ij}^4|) < +\infty$.

Convergence of spike eigenvalues

Consider the sample covariance matrix $S_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^*$, with sample eigenvalues: $\lambda_{n,1} \geq \lambda_{n,2} \geq \cdots \geq \lambda_{n,p}$.

Proposition (Baik and Silverstein - 2006)

Let $s_i = n_1 + \cdots + n_i$ *for* $1 \leq i \leq K$. *Then*

- *For each* $k \in \{1, \ldots, K\}$ *and* $s_{k-1} < j \leq s_k$ *almost surely,*

$$\lambda_{n,j} \longrightarrow \psi(\alpha_k) = \alpha_k + \frac{c\alpha_k}{\alpha_k - 1};$$

- *For all* $1 \leq i \leq L$ *with a prefixed range* L *almost surely,*

$$\lambda_{n,q_0+i} \to b = (1+\sqrt{c})^2.$$

Note. This result has been extended for more general spikes by Bai & Y., Benaych-Georges & Nadakuditi.

b) Estimator of q_0 (number of spikes)

- Based on these results, we observe that when all the spikes are simple, i.e. $n_j \equiv 1$, the spacings

$$\delta_{n,j} = \lambda_{n,j} - \lambda_{n,j+1} \to \begin{cases} r > 0 & \forall j \leq q_0 \\ 0 & \forall j > q_0 \end{cases}$$

- it is possible to detect q_0 form index-number j where $\delta_{n,j}$ becomes small (case of simple spikes). Our estimator is define by

$$\hat{q}_n \quad = \quad \min\{j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n\}, \tag{1}$$

where $(d_n)_n$ is a sequence to be defined and $s > q_0$ is a fixed number.

Consistency of $\hat{q}_n$: case of simple spikes

Assume

- All spikes are different (simple spike case);
- $\sigma^2 = 1$ (if not, take $\delta_{n,j}/\sigma^2$);

and

5. Entries have sub-Gaussian tails: for some positive D, D' we have for all $t \geq$ D',

$$\mathbb{P}(|\mathrm{x_{ij}}| \geq t^D) \leq e^{-t}.$$

Theorem [Passemier & Y. 2011]

Under Assumptions (1)-(5) and in the simple spikes case, if $d_n \to 0$ such that $n^{2/3} d_n \to +\infty$ then

$$\mathbb{P}(\hat{q}_n = q_0) \to 1 \ .$$

Proof (idea)

$$\begin{aligned}\mathbb{P}(\hat{q}_n = q_0) &= 1 - \mathbb{P}\left(\bigcup_{1\le j\le q_0} \{\delta_{n,j} < d_n\} \cup \{\delta_{n,q_0+1} \ge d_n\}\right) \\ &\ge 1 - \sum_{j=1}^{q_0} \mathbb{P}(\delta_{n,j} < d_n) - \underset{(*)}{\mathbb{P}(\delta_{n,q_0+1} \ge d_n)}.\end{aligned}$$

The terms in the sum converge to zero as $d_n \to 0$ and $\delta_{n,j} \to r > 0$. For the last term

$$\begin{aligned}1 - (*) &= \mathbb{P}(n^{2/3}(\lambda_{n,q_0+1} - \lambda_{n,q_0+2}) \le n^{2/3} d_n) \\ &\ge \mathbb{P}\left(\left\{|\mathsf{Y}_{n,1}| \le n^{2/3}\frac{d_n}{2\beta}\right\} \cap \left\{|\mathsf{Y}_{n,2}| \le n^{2/3}\frac{d_n}{2\beta}\right\}\right)\end{aligned}$$

where Y is a tight sequence by the next proposition, and $n^{2/3} d_n / 2\beta \to +\infty$, so $1 - (*) \to 1$.

Proof (an additional important ingredient)

An (partial) extension of Tracy-Widom law in presence of spikes:

Theorem (Benaych-Georges, Guionnet, Maida - 2010)

Under the above assumptions, for all $1 \leq i \leq L$ *with a prefixed range* L

$$\mathrm{Y}_{n,i} = \frac{n^{\frac{2}{3}}}{\beta}(\lambda_{n,q_0+i} - b) = O_{\mathbb{P}}(1)$$

where $\beta = (1+\sqrt{c})(1+\sqrt{c^{-1}})^{\frac{1}{3}}$.

- spacings $\delta_{n,j} \to 0$ from a same spike can also tend to 0;
- Confusion may be possible between these spacings and those from the bulk eigenvalues;
- Hopefully, fluctuations of both type of spacings have different rates:

$$n^{-1/2} \quad \text{v.s.} \quad \simeq n^{-2/3} \, .$$

Theorem (Bai and Y. (2008))

Under Assumptions (1)-(4) (2), the n_k-dimensional real vector

$$\sqrt{n}\{\lambda_{n,j} - \phi(\alpha_k),\, j \in \{s_{k-1}+1, \ldots, s_k\}\}$$

converges weakly to the distribution of the n_k eigenvalues of a Gaussian random matrix whose covariance depend of α_k and c.

[related works are from Baik-Ben-Arous-Pêché, Paul]

Consistency of $\hat{q}_n$: case of multiple spikes

The previous theorem of Bai and Y. implies:

- If $\alpha_j = \alpha_{j+1}$, convergence in $O_{\mathbb{P}}(n^{-1/2})$;
- For unit eigenvalues, faster convergence in $O_{\mathbb{P}}(n^{-2/3})$.

This allows us to use the same estimator provided we use a new threshold d_n.

Theorem (Passemier & Y. (2011))

Under the above assumptions, if

$$d_n = o(n^{-1/2}), \quad \text{and} \quad n^{2/3} d_n \to +\infty,$$

then

$$\mathbb{P}(\hat{q}_n = q_0) \to 1 \ .$$

We decided to use another version of our estimator which performs better

$$\hat{q}_n^* = \min\{j \in \{1, \ldots, s\} : \delta_{n,j+1} < d_n \text{ and } \delta_{n,j+2} < d_n\}$$

Threshold sequence: $d_n = Cn^{-2/3}\sqrt{2\log\log n}$, where C is a constant to be adjusted for each case (Idea: law of the iterated logarithm for $\lambda_{n,j}$, $j \leq q_0$.).

Simulation experiments

- Performance measure: empirical false detection rates over 500 independent replications

$$\mathbb{P}(\tilde{q}_n \neq q_0)$$

- Simulation design:
  - q_0: number of spikes;
  - $(\alpha_i)_{1 \leq i \leq q_0}$: spikes;
  - p: dimension of the vectors;
  - n: sample size;
  - $c = p/n$;
  - $\sigma^2 = 1$ given or to be estimated;
  - C: constant in d_n.

Experimental design

TABLE 1. Summary of parameters used in the simulation experiments. (L: left, R: right)

Fig. No.	Factors	Mod. No.	Factor values	Fixed parameters p, n	c	σ^2	C	Var. par.
1	Different		(α)	$(200, 800)$ $(2000, 500)$	$1/4$ 4	Given	5.5 9	α
2L	Different	A B B	$(6, 5)$ $(10, 5)$ $(10, 5)$		10	Given Estimated	11	n
2R	Different	C D	(1.5) $(2.5, 1.5)$		1	Given	5	n
3	Possibly equal	E F	$(\alpha, \alpha, 5)$ $(\alpha, \alpha, 15)$	$(200, 800)$ $(2000, 500)$	$1/4$ 4	Given	6 9.9	α
4L	Possibly equal	G H H	$(6, 5, 5)$ $(10, 5, 5)$ $(10, 5, 5)$		10	Given Estimated	9.9	n
4R	Possibly equal	I J	$(1.5, 1.5)$ $(2.5, 1.5, 1.5)$		1	Given	5	n
5	Models A and D							
6	Models G and J							
7	No factor	K	No factor		1 10	Given	8 15	n
8L	Models A and G							
8R	Models B and H							
9L	Models C and I, with C automatically chosen							
9R	Models D and J, with C automatically chosen							

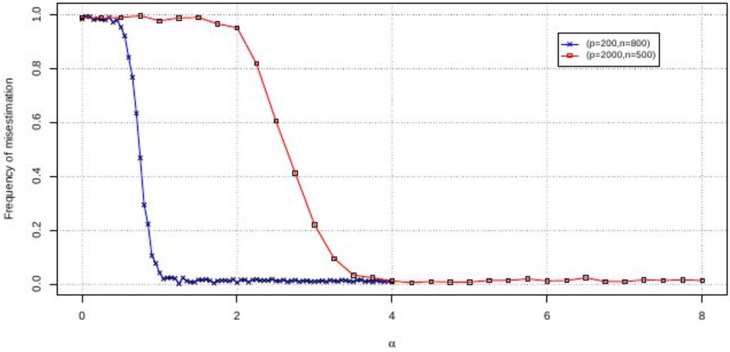


FIGURE 1. Misestimation rates as a function of factor strength for $(p, n) = (200, 800)$ and $(p, n) = (2000, 500)$.

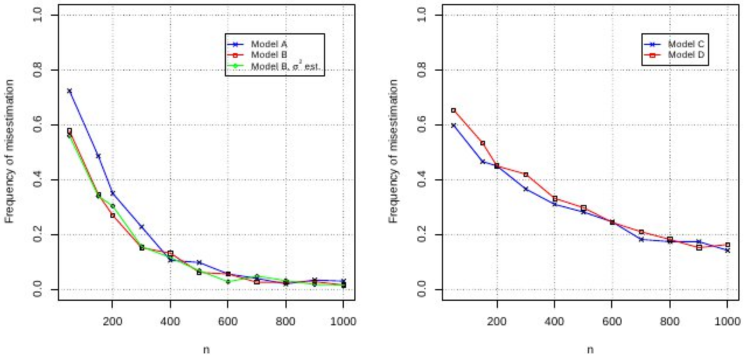


FIGURE 2. Misestimation rates as a function of n for Models A, B (left) and Model C, D (right).

c) Discussions
- Comparison with an estimator by Kritchman and Nadler

In the non-spikes case ($q_0 = 0$), $n\mathrm{S}_n \sim \mathrm{W}_p(\mathrm{I}, n)$. In this case

Proposition (Johnstone - 2001)

$$\mathbb{P}\left(\lambda_{n,1} < \sigma^2 \frac{\beta_{n,p}}{n^{2/3}} s + b\right) \to F_1(s)$$

where F_1 is the Tracy-Widom distribution of order 1 and
$\beta_{n,p} = (1 + \sqrt{p/n})(1 + \sqrt{n/p})^{\frac{1}{3}}$.

To distinguish a spike eigenvalue $\lambda_{n,k}$ from a non-spike one at an asymptotic significance level γ, their idea is to check whether

$$\lambda_{n,k} > \sigma^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b\right)$$

where $s(\gamma)$ verifies $F_1(s(\gamma)) = 1 - \gamma$. Their estimator is

$$\tilde{q}_n = \underset{k}{\operatorname{argmin}} \left(\lambda_{n,k} < \widehat{\sigma}^2 \left(\frac{\beta_{n,p-k}}{n^{2/3}} s(\gamma) + b\right)\right) - 1.$$

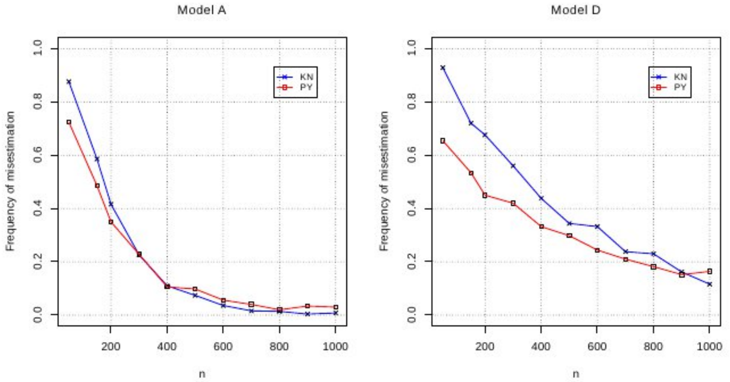


FIGURE 5. Misestimation rates as a function of n for Model A (left) and Model D (right).

c) Discussions
- on the tuning parameter C

- C has been tuned manually in each case ;
- For real applications, need a procedure to choose this constant;
- Idea: use Wishart distributions as a benchmark to calibrate C ;
- consider the gap between two largest eigenvalues: $\tilde{\lambda}_1 - \tilde{\lambda}_2$

Cont'd

- By simulation to get empirical distribution of $\tilde{\lambda}_1 - \tilde{\lambda}_2$;

  500 independent replications.

- compute the upper 5% quantile s:

$$\mathbb{P}(\tilde{\lambda}_1 - \tilde{\lambda}_2 \leq s) \simeq= 0.95 \ .$$

- Define a value

$$\tilde{C} = sn^{2/3} / \sqrt{2 \times \log\log(n)} \ .$$

Results:

TABLE 4. Approximation of the threshold s such that $\mathbb{P}(\tilde{\lambda}_1 - \tilde{\lambda}_2 \leq s) = 0.98$.

(p,n)	(200,200)	(400,400)	(600,600)	(2000,200)	(4000,400)	(7000,700)
Value of s	0.340	0.223	0.170	0.593	0.415	0.306
$\tilde{C}$	6.367	6.398	6.277	11.106	11.906	12.44

Assessment of the automated value $\tilde{C}$ with $c = 10$

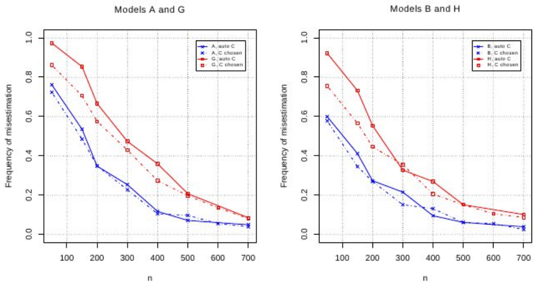
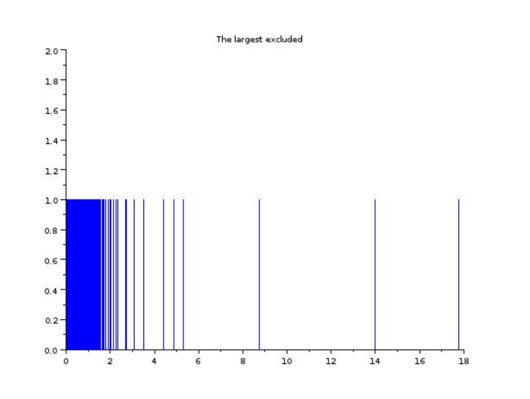


FIGURE 8. Misestimation rates as a function of n for Models A, G (left) and Models B, H (right).

- $\tilde{C} >$ tuned C slightly ;
- Using $\tilde{C} \longrightarrow$ only a small drop of performance ;
- higher error rates in the case of equal factors for moderate sample sizes

Application to S&P stocks data



- Estimated number of factors: $\widehat{q}_0 = 17$;
- Residual variance: $\widehat{\sigma}^2 = 0.3616$.

3) Inference of the bulk spectrum

Estimation of population spectral distribution

$\boxed{\text{Population}}$

$\mathbf{X}$, mean-zero, p-dim
$\mathrm{Cov}(\mathbf{X}) = \Sigma_p$

$\boxed{\text{Sample}}$

$\mathbf{x}_1, \ldots, \mathbf{x}_n$, i.i.d, size n
$S_n = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^* / n$

$\boxed{\text{Large dimensional situations}}$

$$\lim_{n\to\infty} p/n = c > 0$$

$\boxed{\text{PSD } H_p}$

the empirical spectral distribution of Σ_p

$\boxed{\text{ESD } F_n}$

the empirical spectral distribution of S_n.

Problem: Estimate H_p from F_n.

The Marčenko-Pastur equation

- Suppose that
$$p/n \to c > 0, \quad H_p \xrightarrow{w} H,$$
then under suitable conditions, *cf.* Marčenko-Pastur '68, Silverstein '95,
$$F_n \xrightarrow{w} F \ , \quad n \to \infty.$$

- Let $\quad \underline{s}(z) = -(1-c)/z + c\int 1/(x-z)dF(x),$

be the Stieltjes transform of (the companion distribution of) F, then
$$z = -\frac{1}{\underline{s}(z)} + c\int \frac{t}{1+t\underline{s}(z)}dH(t), \quad z \in \mathbb{C}^+,$$
which is called Marčenko-Pastur (MP) equation.

- This gives the inverse map of $\underline{s}(z)$ on $\mathbb{C}\backslash\mathbb{R}$.

Almost all statistical tools for inference of H are based on this equation !!

a). Existing methods for estimation of PSD H

- Inversion of the MP equation:

  1. [El Karoui (2008)], nonparametric, complex field;
  2. [Li et al. (2012)], parametric, real field.

- Methods based on moments of F:

  1. [Rao et al. (2008)], quasi-likelihood;
  2. [Bai et al. (2010)], complete moment method.

- Methods based on moments and contour-integrals:

  1. [Mestre (2008)], eigenvalue splitting condition;
  2. [Yao et al. (2012)], global moment of H;
  3. [Li and Yao (2012)], local moment of H.

Still needs new methods!

However,

- global inversion methods in [El Karoui (2008)] and [Li et al. (2012)] have some implementation issues that are non trivial to overcome;
- other methods are based on moments, but there are situations where these moments can not help to identify model parameters.

Example of a PSD H not identifiable by moments

- H has an inverse cubic density function ([Bouchaud and Potters (2009)])

$$h(t|\alpha) = \frac{b}{(t-a)^3}, \quad t \geq \alpha,$$

where the parameter is $0 \leq \alpha < 1$ is the parameter to be estimated and $a = 2\alpha - 1, \quad b = 2(1-\alpha)^2$.

- Then

$$\int_\alpha xh(x)dx \equiv 1 \ , \quad \int_\alpha x^k h(x)dx = \infty \ , \quad \text{for} \ \ k \geq 2.$$

Moments of H are independent from the parameter α!

b). A generalized expectation based method

Main idea

- Use of general test functions f instead of monomials x^k (moments) ;
- These test functions are usually smaller than the monomials x^k so that

$$T(f) = \int f(x) dH(x)$$

are finite.

In the example above of inverse cubic density, $f(x) = \sin(x)$ has a finite integral:

$$T(f) = b \int_{\alpha}^{\infty} \frac{\sin(x)}{(x-a)^3} dx \ .$$

Generalized expectations and their estimates

- Let f be a analytic function on an open $\mathcal{U} \supset \mathcal{S}_{\mathcal{F}}$, support of F;
- Define a *generalized expectation* $T(f) := \int f(t)dH(t);$
- It will be shown that

$$T(f) = K(c, f) + \frac{1}{2\pi\mathrm{i}c} \oint_{\mathcal{C}} z\underline{s}'(z) f(-1/\underline{s}(z))dz,$$

  where $K(c, f)$ is a constant, independent from H and $\mathcal{C}$ is a contour enclosing S_F.

- With sample eigenvalues, $s(z$ has an empirical estimate

$$\underline{s}_n(z) = -(1 - p/n)/z + (p/n) \int 1/(x - z)dF_n(x)$$

  ,

- Therefore, the above generalized expectation can be estimated by

$$\widehat{T}(f) = K(p/n, f) + \frac{n}{p}\frac{1}{2\pi\mathrm{i}} \oint_{C} z\underline{s}_n'(z) f(-1/\underline{s}_n(z))dz. \qquad (1)$$

Generalized expectation based estimator of H

- Suppose that H belongs to a parametric family:

$$\mathcal{H} = \{H_\theta : \theta \in \Theta \subset \mathbb{R}^q\}.$$

- Construct a q-dim vector of generalized expectations,

$$\gamma = (T(f_j))_{1 \le j \le q} = \left(\int f_j dH_\theta\right) ;$$

such that $g : \theta \mapsto \gamma$ is an one-to-one map on Θ;

- The *generalized expectation estimator* (GEE) of θ is defined to be

$$\widehat{\theta}_n = g^{-1}(\widehat{\gamma}_n),$$

where $\widehat{\gamma}_n = (\widehat{T}(f_j))_{1 \le j \le L_i}$ with elements defined by (1).

c). Asymptotic properties of the GEE estimator

Assumptions:

Assumption (a). $n, p \to \infty$ with $p/n \to c \in (0, \infty)$.

Assumption (b). The sample covariance takes form

$$S_n = \Sigma_p^{1/2} W_n W_n^* \Sigma_p^{1/2} / n,$$

where the entries of $W_n(p \times n)$ are i.i.d. standard real or complex normal variables, and $\Sigma_p^{1/2}$ stands for any Hermitian square root of Σ_p.

Assumption (c). $H_p \xrightarrow{w} H$, a proper probability distribution on $[0, \infty)$. Moreover, the sequence of spectral norms $(\|\Sigma_p\|)$ is bounded.

Asymptotics of $\{\widehat{T}(f_j)\}$'s

Theorem (Li and Y. (2012))

Under the assumptions (a)-(c), *for each* $j = 1, \dots, q$),

1. *the generalized expectation* $T(f_j)$ *can be expressed as*

$$T(f_j) = K(c, f_j) + \frac{1}{2\pi \mathrm{i} c} \oint_{\mathcal{C}} z\underline{s}'(z) f_j(-1/\underline{s}(z)) dz,$$

where the constant $K(c, f_j) = (1 - 1/c) f_j(0)$ *if* $\mathcal{C}$ *encloses 0, and zero otherwise*;

2. *its empirical counterpart* $\widehat{T}(f_j)$ *based on* $\underline{s}_n(z)$ *converges almost surely to* $T(f_j)$;

3. *if in addition, the entries of* W_n $(p \times n)$ *are complex normal, the random vector*

$$n \left[\widehat{T}(f_j) - H_p(f_j) \right]_{1 \le j \le q} \xrightarrow{\mathcal{D}} N_q(0, \Phi),$$

where the centralization term $H_p(f_j)$ *stands for the expectation of* f_j *with respect to* H_p, *where the asymptotic covariances* $\Phi = (\phi_{ij})_{q \times q}$ *are*

$$\phi_{ij} = \frac{-1}{4\pi^2 c^2} \oint_{C} \oint_{C'} f_i(-1/\underline{s}(z_1)) f_j(-1/\underline{s}(z_2)) k(z_1, z_2) dz_1 dz_2,$$

where $k(z_1, z_2) = \underline{s}'(z_1)\underline{s}'(z_2)/(\underline{s}(z_1) - \underline{s}(z_2))^2 - 1/(z_1 - z_2)^2$.

Asymptotics of the GEE estimator $\widehat{\theta}_n$

Theorem (Li and Y. (2012))

In addition to the assumptions (a)-(c)*, suppose that the true value of the parameter* θ_0 *is an inner point of* Θ. *Also, suppose that the function* $g(\theta)$ *is differentiable in a neighborhood of* θ_0 *and the Jacobian matrix* $J(\theta) = \partial g / \partial \theta$ *is invertible at* θ_0. *Then,*

1. *the GEE* $\widehat{\theta}_n$ *is strongly consistent, i.e.*

$$\widehat{\theta}_n \to \theta_0, \quad a.s.,$$

2. *moreover, if in addition, the entries of* W_n $(p \times n)$ *are complex normal, then*

$$n(\widehat{\theta}_n - g^{-1}(\gamma_p)) \xrightarrow{\mathcal{D}} N_q(0, \Gamma(\theta_0)),$$

   where $\gamma_p = (H_p(f_j))_{1 \leq j \leq q}$, *and* $\Gamma(\theta_0) = J^{-1}(\theta_0)\Phi(\theta_0)(J^{-1}(\theta_0))'$ *with* Φ *being defined in Theorem 1.*

d). Application: PSD of S&P 500 stocks covariances

Data analysis:

- Removed the 6 largest eigenvalues (deemed as spike eigenvalues);
- Assume an inverse cubic density for PSD H associated to the 482 bulk eigenvalues, that is,

$$h(t|\alpha) = \frac{b}{(t-a)^3}, \quad t \geq \alpha \ ,$$

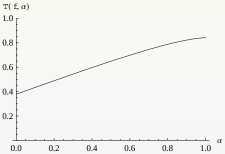
  where $0 < \alpha < 1, \ b = 2(1-\alpha)^2$ and $a = 2\alpha - 1$;
- Moments-based methods fail, LEE may work!

Application to S&P 500 stocks data

- Consider

$$f(z) = \sin(z), \quad T(f,\alpha) = \int \sin(t) h(t|\alpha) dt;$$

- $T(f,\alpha)$ is increasing with respect to α,



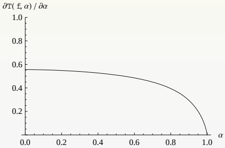


Figure: Curves of $T(f,\alpha)$ (left) and $\partial T(f,\alpha)/\partial\alpha$ (right).

Results on S&P 500 stocks data

- GEE: $\widehat{T}(f,\alpha) = 0.5546$, $\widehat{\alpha} = 0.3205$;
- LSE: $\widehat{\alpha}' = 0.4384$ (see [Li et al. (2012)]);
- Denote by f_α the density function of LSD F with respect to $H(\alpha)$. Compute a kernel density estimate $\widehat{f}_{ker}$ from the 482 bulk eigenvalues (Gaussian kernel, bandwidth $h = 0.01$).
- Consider $d(\alpha) = L^2(f_\alpha, \widehat{f}_{ker})$, then $d(\widehat{\alpha}) = 0.2047$, $d(\widehat{\alpha}') = 0.2863$.

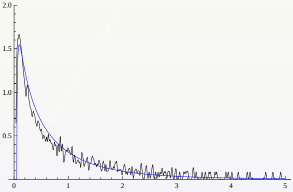


Figure: $\widehat{f}_{ker}$ (plain black), $f_{\widehat{\alpha}}$ (left, blue), and $f_{\widehat{\alpha}'}$ (right, blue).

- GEE yields a significantly better fit to the density of bulk eigenvalues.

Thank you !

Bai, Z. D., Chen, J. Q. and Yao, J. F. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.* **52** 423–437.

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.*, 32, 553–605.

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral analysis of large dimensional random matrices*, 2nd ed. Springer, New York.

Bouchaud, J. P. and Potters, M. (2009). Financial applications of random matrix theory: A short review. *ArXiv:0910.1205v1*.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.*, 24, 508–539.

El Karoui, N. (2008) Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790.

Li, W. M., Chen, J. Q., Qin, Y. L., Yao, J. F. and Bai, Z. D. (2011) Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *Submitted.*

Li, W. M. and Yao, J. F. (2011) A local moments estimation of the spectrum of a large dimensional covariance matrix. *Submitted*.

Mestre, X. (2008) Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory* **54**, 5113–5129.

Rao, N. R., Mingo, J. A., Speicher, R. and Edelman, A. (2008) Statistical eigen-inference from large Wishart matrices. *Ann. Statist.* **36** 2850–2885.

Yao, J. F., Kammoun, A., and Najim, J. (2012) Estimation of the covariance matrix of large dimensional data. *ArXiv:1201.4672v1.*