ASYMPTOTIC THEORY FOR LARGE SAMPLE COVARIANCE MATRICES

JOHANNES HEINY, RICHARD A. DAVIS, THOMAS MIKOSCH

UNIVERSITY OF COPENHAGEN

KØBENHAVNS UNIVERSITET

ABSTRACT

In risk management an appropriate assessment of the dependence structure of multivariate data plays a crucial role for the trustworthiness of the obtained results. The case of *heavy-tailed components* is of particular interest.
We consider asymptotic properties of sample covariance matrices for such time series, where both the dimension and the sample size tend to infinity simultaneously.

KNOWN RESULTS

If the rows of $\boldsymbol{X}$ are independent and identically distributed strictly stationary ergodic time series, then for fixed p we have $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}' \xrightarrow{\text{a.s.}} \boldsymbol{I}_p$.
In particular, if $\boldsymbol{X}$ has iid standard normal entries Johnstone (2001) showed that for $p, n \to \infty$ with $p/n \to \gamma > 0$,

$$\frac{\sqrt{n}+\sqrt{p}}{(1/\sqrt{n}+1/\sqrt{p})^{1/3}}\Big(\frac{\lambda_{(1)}}{(\sqrt{n}+\sqrt{p})^2}-1\Big) \xrightarrow{d} \text{TW},$$

a Tracy-Widom distribution.
Let us now assume that the entries of $\boldsymbol{X}$ are still iid but with *infinite fourth moment* (heavy tails). Since $\limsup \lambda_{(1)}/n = \infty$ a.s. a much stronger normalization of $\boldsymbol{X}\boldsymbol{X}'$ is required.

OUR MODEL

Suppose $\boldsymbol{X} = (X_{it})_{i=1,\dots,p;t=1,\dots,n}$ with

$$X_{it} = \sum_{l=0}^{\infty}\sum_{k=0}^{\infty} h_{kl} Z_{i-k,t-l}$$

and **regularly varying** iid noise (Z_{it}) with index $\alpha \in (0,4)$ (infinite fourth moment), i.e. there exists a normalizing sequence (a_n) such that

$$n\mathbb{P}(|Z| > a_n x) \to x^{-\alpha}, \quad \text{as } n \to \infty \text{ for } x > 0,$$

and a tail balance condition holds. If Z is regularly varying with index α, then moments above the αth do not exist.
Moreover we impose a summability condition on the double array of real numbers (h_{kl}) and rather technical growth conditions on $p = p_n \to \infty$.

SETUP & OBJECTIVE

Data matrix: a $p \times n$ matrix $\boldsymbol{X}$ consisting of n observations of a p-dimensional time series, i.e.

$$\boldsymbol{X} = (X_{it})_{i=1,\dots,p;t=1,\dots,n}.$$

We are interested in the non-normalized $p\times p$ *sample covariance matrix* $\boldsymbol{X}\boldsymbol{X}'$ and its *ordered eigenvalues*

$$\lambda_{(1)} \geq \lambda_{(2)} \geq \dots \geq \lambda_{(p)}.$$

MAIN RESULT

The order statistics $D_{(i)}$ of the iid sequence $D_s = \sum_{t=1}^{n} Z_{st}^2$ and the ordered eigenvalues $v_{(j)}$ of the matrix M given by $M_{ij} = \sum_{\ell=0}^{\infty} h_{i\ell}h_{j\ell}$ play a key role in determining the asymptotic properties of the ordered eigenvalues $\lambda_{(i)}$. Let $k^2 = o(p)$ be an integer sequence.

Theorem. If $\alpha \in (0,2)$, then

$$a_{np}^{-2} \max_{i=1,\dots,p} |\lambda_{(i)} - \delta_{(i)}| \xrightarrow{\mathbb{P}} 0, \quad n \to \infty,$$

where $\delta_{(1)} \geq \dots \geq \delta_{(p)}$ are the ordered values of the set $\{D_{(i)}v_{(j)} : i \leq k; j \geq 1\}$.

POINT PROCESS CONVERGENCE

Let (E_i) be iid standard exponential random variables and $\Gamma_i = E_1 + \ldots + E_i$. Then we have the point process convergence

$$\sum_{i=1}^{p} \varepsilon_{a_{np}^{-2}\lambda_i} \xrightarrow{d} \sum_{i=1}^{\infty}\sum_{j=1}^{r} \varepsilon_{\Gamma_i^{-2/\alpha}v_j}. \qquad (4)$$

An application of (4) then yields for every fixed integer $k \geq 1$,

$$a_{np}^{-2}\big(\lambda_{(1)},\dots,\lambda_{(k)}\big) \xrightarrow{d} \big(d_{(1)},\dots,d_{(k)}\big),$$

where $d_{(1)} \geq \dots \geq d_{(k)}$ are the k largest ordered values of the set $\{\Gamma_i^{-2/\alpha}v_j : i, j \geq 1\}$. In particular we find

$$d_{(1)} = v_1\Gamma_1^{-2/\alpha} \text{ and } d_{(2)} = v_2\Gamma_1^{-2/\alpha} \vee v_1\Gamma_2^{-2/\alpha}. \qquad (5)$$

EXAMPLE

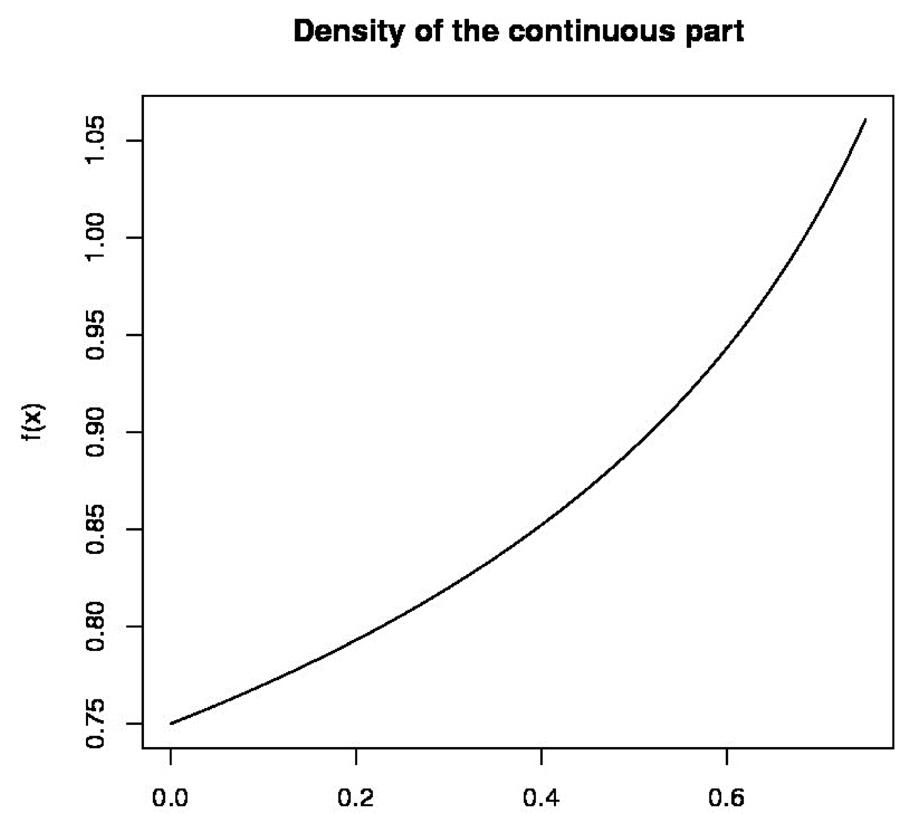

Figure 1: The density of the continuous part of Y defined in (2) with $\alpha = 1.5$.

Assume that $\alpha \in (0,2)$ and

$$X_{it} = Z_{it} + Z_{i,t-1} - 2(Z_{i-1,t} - Z_{i-1,t-1}). \qquad (1)$$

The matrix M has rank 2 and the non-negative eigenvalues $v_1 = 8$ and $v_2 = 2$. The limit point process in (4) is

$$\sum_{i=1}^{\infty} \varepsilon_{8\Gamma_i^{-2/\alpha}} + \sum_{i=1}^{\infty} \varepsilon_{2\Gamma_i^{-2/\alpha}}.$$

By (5) we get

$$a_{np}^{-2}\lambda_{(2)} \xrightarrow{d} 2\Gamma_1^{-2/\alpha} \vee 8\Gamma_2^{-2/\alpha}.$$

Since Γ_1/Γ_2 has a standard uniform distribution, we can easily compute

$$\mathbb{P}(2\Gamma_1^{-2/\alpha} > 8\Gamma_2^{-2/\alpha}) = 2^{-\alpha} \in (1/4, 1).$$

The self-normalized spectral gap

$$\frac{\lambda_{(1)} - \lambda_{(2)}}{\lambda_{(1)}}$$

converges in distribution to a random variable which has the same distribution as

$$Y := 3/4 I_{\{U<2^{-\alpha}\}} + \big(1 - U^{2/\alpha}\big) I_{\{U>2^{-\alpha}\}}, \qquad (2)$$

where U is standard uniformly distributed. Y has an atom at $3/4$ with point mass $2^{-\alpha}$. The ratio of the two largest eigenvalues is of special interest. In the case of independent rows it was shown that $\lambda_{(2)}/\lambda_{(1)} \to U^{\alpha/2}$ in distribution. In our model, however, the rows are dependent and the limit takes the form

$$c^{\alpha/2} I_{\{U<c\}} + U^{\alpha/2} I_{\{U>c\}}$$

for a non-negative constant c. To confirm this limit structure we simulate the ratio $(\lambda_{(2)}/\lambda_{(1)})^{2/\alpha}$ from the model (1) for $\alpha = 1.5$. The theoretical limit variable is

$$(1 - Y)^{2/\alpha} = 0.35 I_{\{U<0.35\}} + U_{\{U>0.35\}}. \qquad (3)$$

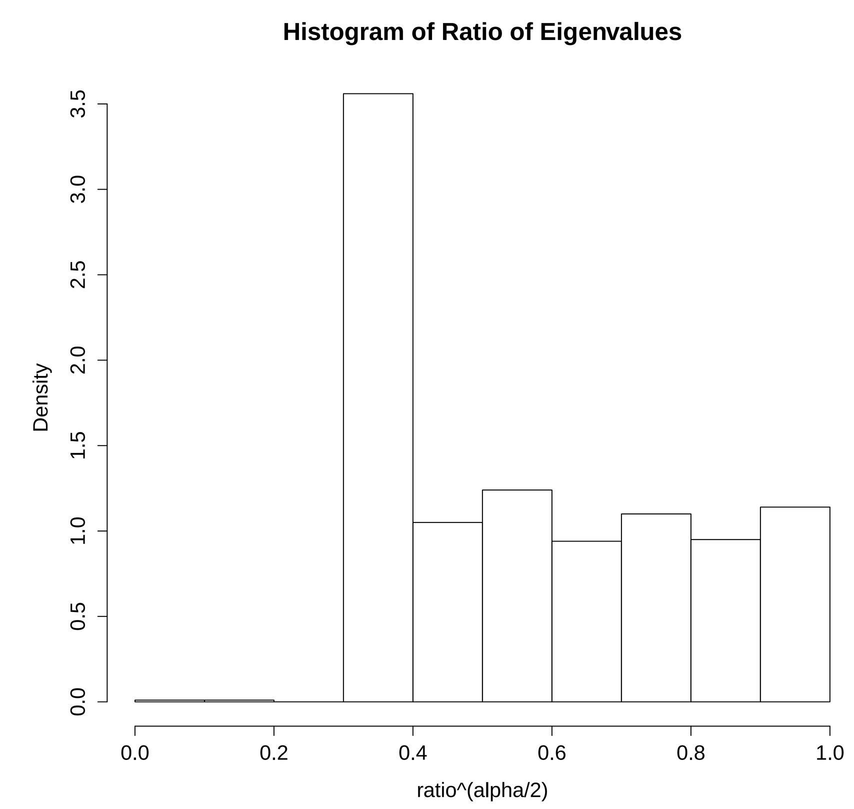

Figure 2: The histogram of $\big(\lambda_{(2)}/\lambda_{(1)}\big)^{2/\alpha}$ based on 1000 replications from the model (1) with noise given by a t-distribution with $\alpha = 1.5$ degrees of freedom, $n = 1000$ and $p = 200$.

A histogram based on realizations of the true limit variable (3) would look very similar.

REFERENCES, FUTURE RESEARCH & CONTACT INFORMATION

1. Davis, Mikosch, Pfaffel. *Asymptotic theory for the sample covariance matrix of a heavy-tailed multivariate time series.* Working paper.
2. Johnstone (2001). *On the distribution of the largest eigenvalue in principal component analysis.* Ann. Statist. **29** (2), 295–327.

- Autocovariance matrix.
- Centering in the case $\alpha \in (2,4)$.
- Determinants and matrix decompositions.
- Other non-linear structures of X_{it}.

Johannes Heiny
`johannes.heiny@ math.ku.dk`