Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications ▮ (▮▮▮▮) ▮▮▮–▮▮▮

stochastic processes and their applications

www.elsevier.com/locate/spa

Almost sure convergence of the largest and smallest eigenvalues of high-dimensional sample correlation matrices☆

Johannes Heiny[a,*], Thomas Mikosch[b]

[a] *Department of Mathematics, Aarhus University, Ny Munkegade 118, DK-8000 Aarhus C, Denmark*
[b] *Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark*

Received 20 December 2016; received in revised form 19 September 2017; accepted 6 October 2017
Available online xxxx

Abstract

In this paper, we show that the largest and smallest eigenvalues of a sample correlation matrix stemming from n independent observations of a p-dimensional time series with iid components converge almost surely to $(1+\sqrt{\gamma})^2$ and $(1-\sqrt{\gamma})^2$, respectively, as $n \to \infty$, if $p/n \to \gamma \in (0,1]$ and the truncated variance of the entry distribution is "almost slowly varying", a condition we describe via moment properties of self-normalized sums. Moreover, the empirical spectral distributions of these sample correlation matrices converge weakly, with probability 1, to the Marčenko–Pastur law, which extends a result in Bai and Zhou (2008). We compare the behavior of the eigenvalues of the sample covariance and sample correlation matrices and argue that the latter seems more robust, in particular in the case of infinite fourth moment. We briefly address some practical issues for the estimation of extreme eigenvalues in a simulation study.

In our proofs we use the method of moments combined with a Path-Shortening Algorithm, which efficiently uses the structure of sample correlation matrices, to calculate precise bounds for matrix norms. We believe that this new approach could be of further use in random matrix theory.

☆ Johannes Heiny's and Thomas Mikosch's research is partly supported by the Danish Research Council Grant DFF-4002-00435 "Large random matrices with heavy tails and dependence". Johannes Heiny thanks the Villum Foundation for the grant 11745 "Ambit fields: Probabilistic properties and statistical inference".

* Corresponding author.
E-mail addresses: johannes.heiny@math.ku.dk (J. Heiny), mikosch@math.ku.dk (T. Mikosch).
URL: http://www.math.ku.dk/∼mikosch (T. Mikosch).

https://doi.org/10.1016/j.spa.2017.10.002

MSC: primary 60B20; secondary 60F05; 60F10; 60G10; 60G55; 60G70

Keywords: Sample correlation matrix; Infinite fourth moment; Largest eigenvalue; Smallest eigenvalue; Spectral distribution; Sample covariance matrix; Self-normalization; Regular variation; Combinatorics

1. Introduction and notation

In modern statistical analyses one is often faced with large data sets where both the dimension of the observations and the sample size are large. The dramatic increase and improvement of computing power and data collection devices have triggered the necessity to study and interpret the sometimes overwhelming amounts of data in an efficient and tractable way. Huge data sets arise naturally in wireless communication, finance, natural sciences and genetic engineering. For such data one commonly studies the dependence structure via covariances and correlations which can be estimated by their sample analogs. *Principal component analysis*, for example, uses an orthogonal transformation of the data such that only a few of the resulting vectors explain most of the variation in the data. The empirical variances of these so-called *principal component vectors* are the largest eigenvalues of the *sample covariance or correlation matrix*.

Throughout this paper we consider the $p \times n$ data matrix

$$\mathbf{X} = \mathbf{X}_n = (X_{it})_{i=1,\ldots,p;t=1,\ldots,n}$$

of identically distributed entries (X_{it}) with generic element X, where we assume $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$ if the first and second moments of X are finite, respectively. A column of $\mathbf{X}$ represents an observation of a p-dimensional time series.

Random matrix theory provides a great variety of results on the ordered eigenvalues

$$\lambda_{(1)} \geq \cdots \geq \lambda_{(p)}, \tag{1.1}$$

of the (non-normalized) *sample covariance matrix* $\mathbf{X}\mathbf{X}'$. Here we will only discuss the case $p = p_n \to \infty$ and, unless stated otherwise, we assume the growth condition

$$\lim_{n\to\infty} \frac{p_n}{n} = \gamma \in (0, 1]. \tag{G_γ}$$

For the finite p case, we refer to [3,32,26]. When studying the asymptotic properties of estimators under (G_γ) one often obtains results that dramatically differ from the standard p fixed, $n \to \infty$ case, in which the spectrum of $(n^{-1}\mathbf{X}\mathbf{X}')$ converges to its population covariance spectrum. In 1967, Marčenko and Pastur [30] observed that even in the case of iid entries (X_{it}) with $\mathbb{E}[X^2] = 1$ the eigenvalues $(\lambda_{(i)}/n)$ do not concentrate around 1. For more examples, see [6, Chapter 1] and [21]. Typical applications where (G_γ) seems reasonable are discussed in [28,19].

In comparison with $(\lambda_{(i)})$, much less is known about the ordered eigenvalues

$$\mu_{(1)} \geq \cdots \geq \mu_{(p)}$$

of the *sample correlation matrix* $\mathbf{R} = \mathbf{Y}\mathbf{Y}'$ with entries

$$R_{ij} = \sum_{t=1}^{n} \frac{X_{it}X_{jt}}{\sqrt{D_i}\sqrt{D_j}} = \sum_{t=1}^{n} Y_{it}Y_{jt}, \quad i, j = 1, \ldots, p. \tag{1.2}$$

Please cite this article in press as: J. Heiny, T. Mikosch, Almost sure convergence of the largest and smallest eigenvalues of high-dimensional sample correlation matrices, Stochastic Processes and their Applications (2017), https://doi.org/10.1016/j.spa.2017.10.002.