Journal of Statistical Planning and Inference 150 (2014) 1–29

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Review

Random matrix theory in statistics: A review

Debashis Paul*, Alexander Aue

Department of Statistics, University of California, Davis, United States

A R T I C L E   I N F O

Article history:
Received 1 March 2013
Received in revised form
6 September 2013
Accepted 10 September 2013
Available online 27 September 2013

Keywords:
MANOVA
Marčenko–Pastur law
Principal component analysis
Stieltjes transform
Wishart matrix
Tracy–Widom law

A B S T R A C T

We give an overview of random matrix theory (RMT) with the objective of highlighting the results and concepts that have a growing impact in the formulation and inference of statistical models and methodologies. This paper focuses on a number of application areas especially within the field of high-dimensional statistics and describes how the development of the theory and practice in high-dimensional statistical inference has been influenced by the corresponding developments in the field of RMT.

Contents

* Corresponding author. Tel.: +1 530 564 0513.
E-mail addresses: debashis@wald.ucdavis.edu, debpaul@ucdavis.edu (D. Paul), alexaue@wald.ucdavis.edu (A. Aue).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.09.005

1. Introduction

Statistics has entered into an age where an increasingly larger volume of more complex data is being generated, often through automated measuring devices, in a wide array of disciplines such as genomics, atmospheric sciences, communications, biomedical imaging, economics and many others. The representation of such data in any nominal coordinate system often leads to so-called high-dimensional data that are frequently associated with phenomena transcending the boundary of classical multivariate statistical analysis. The continued growth of these new data sources has given rise to the incorporation of different mathematical tools into the realms of statistical analysis, which include convex analysis, Riemannian geometry and combinatorics. Random matrix theory has emerged as a particularly useful framework for posing many theoretical questions associated with the analysis of high-dimensional multivariate data.

In this paper, we mainly focus on several application areas of random matrix theory (RMT) in statistics. These include problems in dimension reduction, hypothesis testing, clustering, regression analysis and covariance estimation. We also briefly describe the important role played by RMT in enabling certain theoretical analyses in wireless communications and econometrics. Different themes emerging from these problems have in turn led to further investigation of some classical RMT phenomena. Among these, the notion of *universality* has profound implications in the context of high-dimensional data analysis in terms of the applicability of many statistical techniques beyond the classical framework built upon the multivariate Gaussian distribution. With these perspectives, the treatment of the topics will focus on those aspects of RMT relevant to the statistical questions. Thus, the topics in RMT that receive most attention in this paper are those related to the behavior of the bulk spectrum, i.e., the empirical spectral distribution, and the behavior of the edge of the spectrum, i.e., the extreme eigenvalues, of random matrices. Also, the sample covariance matrix being a dominant object of study in most of the multivariate analyses, much of the paper is devoted to the study of its spectral behavior. For more detailed accounts of these topics, the reader may refer to Anderson et al. (2009), Bai and Silverstein (2009) and Pastur and Shcherbina (2011). More complete and self-containing treatments of a number of "core" topics of RMT not covered in this paper, including the Riemann–Hilbert approach to the asymptotics of orthogonal polynomials and random matrices, the distribution of spacings and correlation functions of eigenvalues and their connections with determinantal point processes, and the role of the eigenvalue statistics in physics, can be found in the monographs (Akemann et al., 2011; Anderson et al., 2009; Deift, 2000; Deift and Gioev, 2009; Forrester, 2010; Guionnet, 2009; Mehta, 2004; Tao, 2012), and the survey articles (Diaconis, 2003; Soshnikov, 2000). The connection between RMT and free probability theory, another significant topic not discussed here, is explored in detail in Anderson et al. (2009), Hiai and Petz (2000), Mingo and Specicher (2006), Nica and Speicher (2006) and Edelman and Rao (2005). Finally, wireless communications and finance are two areas beyond physics and statistics where tools and concepts from RMT have been successfully applied and thus we give a brief overview of these topics in Section 4. Applications of RMT in wireless communications are the focus of Coulliet and Debbah (2011) and Tulino and Verdú (2004), while for a detailed look at applications in finance one may refer to Bouchaud et al. (2003) and Bouchaud and Potters (2009). A survey of some of the statistical topics covered in this review can be found in Johnstone (2007).

We now give a brief outline of this paper. There are two different ways in which RMT has impacted modern statistical procedures. On one hand, most of the mathematical treatment of RMT have focused on matrices with high degree of independence in the entries, which one may refer to as "unstructured" random matrices. The results from the corresponding theory have been used primarily in the context of hypothesis testing where the null hypothesis corresponds to the absence of any directionality, or signal component, in the data. On the other hand, in high-dimensional statistics, we are primarily interested in problems where there are lower dimensional structures buried under random noise. An effective treatment of the latter problem often requires going beyond the realms of the classical RMT framework and into the domain of statistical regularization schemes. Keeping these perspectives in mind, we devote Sections 2 and 3 to the motivations and theoretical developments in RMT, while focusing on the statistical applications of RMT in Section 4. Finally, in Section 5, we focus on modern statistical regularization schemes based on various forms of sparse structures for dealing with high-dimensional statistical problems.