An Introduction to Wishart Matrix Moments

Other titles in Foundations and Trends® in Machine Learning

Non-convex Optimization for Machine Learningy
Prateek Jain and Purushottam Ka
ISBN: 978-1-68083-368-3

Kernel Mean Embedding of Distributions: A Review and Beyond
Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Scholkopf
ISBN: 978-1-68083-288-4

Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 1 Low-Rank Tensor Decompositions
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-222-8

Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 2 Applications and Future Perspectives
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-276-1

Patterns of Scalable Bayesian Inference
Elaine Angelino, Matthew James Johnson and Ryan P. Adams
ISBN: 978-1-68083-218-1

Generalized Low Rank Models
Madeleine Udell, Corinne Horn, Reza Zadeh and Stephen Boyd
ISBN: 978-1-68083-140-5

An Introduction to Wishart Matrix Moments

Adrian N. Bishop
University of Technology Sydney (UTS) and CSIRO, Australia
adrian.bishop@uts.edu.au

Pierre Del Moral
INRIA, Bordeaux Research Center, France
pierre.del-moral@inria.fr

Angèle Niclas
École Normale Supérieure de Lyon, France

the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. N. Bishop, P. Del Moral and A. Niclas. *An Introduction to Wishart Matrix Moments*. Foundations and Trends® in Machine Learning, vol. 11, no. 2, pp. 97–218, 2018.

ISBN: 978-1-68083-507-6

Foundations and Trends® in Machine Learning

Volume 11, Issue 2, 2018

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2018, Volume 11, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

An Introduction to Wishart Matrix Moments

Adrian N. Bishop[1], Pierre Del Moral[2] and Angèle Niclas[3]

[1] *University of Technology Sydney (UTS) and CSIRO, Australia; adrian.bishop@uts.edu.au*

[2] *INRIA, Bordeaux Research Center, France; pierre.del-moral@inria.fr*

[3] *École Normale Supérieure de Lyon, France*

ABSTRACT

These lecture notes provide a comprehensive, self-contained introduction to the analysis of Wishart matrix moments. This study may act as an introduction to some particular aspects of random matrix theory, or as a self-contained exposition of Wishart matrix moments.

Random matrix theory plays a central role in statistical physics, computational mathematics and engineering sciences, including data assimilation, signal processing, combinatorial optimization, compressed sensing, econometrics and mathematical finance, among numerous others. The mathematical foundations of the theory of random matrices lies at the intersection of combinatorics, non-commutative algebra, geometry, multivariate functional and spectral analysis, and of course statistics and probability theory. As a result, most of the classical topics in random matrix theory are technical, and mathematically difficult to penetrate for non-experts and regular users and practitioners.

The technical aim of these notes is to review and extend some important results in random matrix theory in the specific

Adrian N. Bishop, Pierre Del Moral and Angèle Niclas (2018), "An Introduction to Wishart Matrix Moments", Foundations and Trends® in Machine Learning: Vol. 11, No. 2, pp 97–218. DOI: 10.1561/2200000072.

context of real random Wishart matrices. This special class of Gaussian-type sample covariance matrix plays an important role in multivariate analysis and in statistical theory. We derive non-asymptotic formulae for the full matrix moments of real valued Wishart random matrices. As a corollary, we derive and extend a number of spectral and trace-type results for the case of non-isotropic Wishart random matrices. We also derive the full matrix moment analogues of some classic spectral and trace-type moment results. For example, we derive semi-circle and Marchencko–Pastur-type laws in the non-isotropic and full matrix cases. Laplace matrix transforms and matrix moment estimates are also studied, along with new spectral and trace concentration-type inequalities.

1

Introduction

Let X be a centered Gaussian random column vector with covariance matrix P on $\mathbb{R}^r$, for some dimension parameter $r \geq 1$. The rescaled sample covariance matrix associated with $(N+1)$ independent copies X_i of X is given by the random matrix

$$P_N = \frac{1}{N} \sum_{1 \leq i \leq N+1} \left(X_i - m^N\right) \left(X_i - m^N\right)'$$

with the sample mean

$$m^N := \frac{1}{N+1} \sum_{1 \leq i \leq N+1} X_i$$

Here, $(.)'$ denotes the transpose operator. The random matrix P_N has a Wishart distribution with N degrees of freedom and covariance matrix $N^{-1}P$ (a.k.a. the scale matrix). When $N \geq r$,the distribution of the Wishart matrix P_N on the cone of symmetric positive definite matrices is defined by

$$\begin{aligned}
&\text{Probability}(P_N \in dQ) \\
&\quad = \frac{\det(Q)^{(N-r-1)/2}}{2^{Nr/2}\Gamma_r(N/2)\det(P/N)^{N/2}} \exp\left[-\tfrac{1}{2}\mathrm{Tr}\left((P/N)^{-1}Q\right)\right]\gamma(dQ)
\end{aligned}$$

where $\det(Q)$ denotes the determinant of Q and $\gamma(dQ)$ is the Lebesgue measure on the cone of symmetric positive definite matrices, and Γ_r is the multivariate gamma function

$$\Gamma_r(z) = \pi^{r(r-1)/4} \prod_{1 \leq k \leq r} \Gamma\left(z - \frac{k-1}{2}\right)$$

We also have the equivalent formulations

$$P_N \stackrel{law}{=} N^{-1} \sum_{1 \leq i \leq N} \mathbb{X}_i = \mathcal{X}\mathcal{X}'$$

with the $(r \times N)$-random matrix $\mathcal{X}$ defined by

$$\mathcal{X} = \frac{1}{\sqrt{N}} [X_1, \dots, X_N]$$

In the above display, $\mathbb{X}_i$ stand for N independent copies of the rank one random matrix $\mathbb{X} = XX'$. The superscript $(.)'$ denotes the transposition operation.

Random matrices, sample covariance matrices, and more specifically Wishart random matrices, play a role in finance and statistics, physics, and engineering sciences. Their interpretation depends on the application model motivating their study.

For example, in Bayesian inference, Wishart matrices often represent the prior precision matrix of multivariate Gaussian data sets. In this context, the posterior distribution of the random covariance given the multivariate-normal vector is again a Wishart distribution with a scale matrix that depends on the measurements. In other words, Wishart distributions are conjugate priors of the inverse covariance-matrix of a multivariate normal random vector [8, 64].

In multivariate analysis and machine learning, the vectors X_i may represent some statistical data such as image, curves and text data. In this case, P may be defined in terms of some covariance function as in Gaussian processes [72]. As its name indicates, the sample covariance matrix P_N attempts to capture the shape of the data; such as the spread around their sample mean as well as the sample correlation between the features dimensions. Principal component analysis and related techniques amount to finding the eigenvalues and the corresponding eigenvectors of sample covariance matrices. The largest eigenvalues represents the dimensions with the strongest correlation in the data set. Expressing the data on the eigenvectors associated with the largest eigenvalues is often used to compress high dimensional data. For a more thorough discussion on this subject we refer to the articles [2, 7, 46, 69], as well as the monographs [8, 72, 63] and the references therein.

In the context of multiple-input multiple-output systems, more general random matrices may be related to the channel gain matrix [77, 99]. Similarly, the covariance matrix in Gaussian process-based inference may be considered a random matrix defined by the particular covariance structure [72]. In data assimilation problems and filtering theory, non-independent sample covariance matrices arise as the control gain in ensemble (Kalman-type) filters; see e.g. [11, 23, 9] and the references therein. Similar (non-independent) sample covariance matrices may be computed with the particles in classical Markov Chain Monte Carlo and sequential Monte Carlo methods [21]; and in this case often represent the uncertainty in an estimation theoretic sense. In finance, sample covariance matrices arise in risk management and asset allocation; e.g. random matrices may represent the correlated fluctuations of assets [57, 13, 27].

Because of their practical importance, we may illustrate the above specific model via the so-called Wishart process. Consider a time-varying linear-Gaussian diffusion of the following form,

$$dX(t) = A(t)\, X(t)\ dt + R(t)^{1/2}\, dB(t) \tag{1.1}$$

where $B(t)$ is an r-dimensional Brownian motion, X_0 is a r-dimensional Gaussian random variable with mean and variance $(\mathbb{E}(X_0), P_0)$, independent of $B(t)$, and $A(t) \in \mathbb{R}^{r\times r}$, and $R(t) > 0$ is a positive definite symmetric matrix. The covariance matrices

$$P(t) = \mathbb{E}\left([X(t) - \mathbb{E}(X(t))]\, [X(t) - \mathbb{E}(X(t))]' \right)$$

satisfy the (linear) matrix-valued differential equation,

$$\partial_t P(t) = A(t)\, P(t) + P(t)\, A(t)' + R(t)$$

The solution of the preceding equation is given easily via the transition/fundamental matrix defined by $A(t)$. More precisely, the solution of the above equation is given by the formula

$$P(t) = e^{\oint_0^t A(s)ds}\ P(0)\ \left[e^{\oint_0^t A(s)ds}\right]' \\ + \int_0^t\ e^{\oint_s^t A(u)du}\ R(s)\ \left[e^{\oint_s^t A(u)du}\right]'\ ds$$

In the above display, $\mathcal{E}_{s,t} := e^{\oint_s^t A(u)du}$ denotes the matrix exponential semigroup, or the transition matrix, defined by

$$\partial_t\, \mathcal{E}_{s,t} = A_t\, \mathcal{E}_{s,t} \quad \text{and} \quad \partial_s\, \mathcal{E}_{s,t} = -\mathcal{E}_{s,t}\, A_s \quad \text{with} \quad \mathcal{E}_{s,s} = I$$

For time homogeneous models $(A(t), R(t)) = (A, R)$ the above formula reduces to

$$P(t) = e^{tA} P(0) + \int_0^t\ e^{(t-s)A}\ R\ e^{(t-s)A'}\ ds$$

The rescaled sample covariance matrices associated with $(N+1)$ independent copies $(X_i(t))_{1\leq i\leq N+1}$ of the process $X(t)$ are defined by

$$P_N(t) := \frac{1}{N} \sum_{1\leq i\leq N+1} \left[X_i(t) - m^N(t)\right] \left[X_i(t) - m^N(t)\right]'$$

with the sample mean

$$m^N(t) := \frac{1}{N+1} \sum_{1\leq i\leq N+1} X_i(t)$$

Up to a change of probability space, the process $P_N(t)$ satisfies the matrix diffusion equation

$$dP_N(t) = A(t)\, P_N(t) + P_N(t)\, A(t)' + R(t) + \frac{1}{\sqrt{N}}\ M_N(t)$$

with the matrix-valued martingale

$$dM_N(t) = P_N(t)^{1/2}\ d\mathcal{W}(t)\ R(t)^{1/2} + R(t)^{1/2}\ d\mathcal{W}(t)\ P_N(t)^{1/2}$$

where $\mathcal{W}_t$ denotes an $(r\times r)$-matrix with independent Brownian entries. The above diffusion coincides with the Wishart process considered in [15]. When $r = 1$ this Wishart model coincides with the Cox–Ingersoll–Ross process (a.k.a. squared Bessel process) introduced in [19]. For a more detailed discussion on Wishart processes and related affine diffusions, we refer to the articles [20, 40, 58], and the references therein.

The preceding exposition is by no means exhaustive of applications of random, and more specifically Wishart, matrices and we point to [26, 30, 59, 89, 97, 98] for further applications and motivators. The typical technical questions arising in practice revolve around the calculation of the spectrum distribution, and the corresponding eigenvector distribution of these random matrices.

The analysis of Wishart matrices started in 1928 with the pioneering work of J. Wishart [100]. Since this, the theory of random matrices has been a fruitful contact point between statistics, pure and applied probability, combinatorics, non-commutative algebra, as well as differential geometry and functional analysis.

The joint distribution of the eigenvalues of real valued Wishart matrices is only known for full rank and isotropic models; that is when the sample size is greater than the dimension, and the covariance matrix $P \propto I$ is proportional to the identity matrix I; see for instance [34, 62]. In this situation, the matrix of random eigenvectors is uniformly distributed on the manifold of unitary matrices equipped with the Haar measure. In this context, marginal distributions for these uncorrelated models can also be computed in a tractable form only for the smallest and the largest eigenvalues. Sophisticated integral formulae for the marginal distribution of intermediate eigenvalues are provided by Zanella–Chiani–Win [104]. Upper bounds on the marginal distribution of the ordered eigenvalues are given in [67].

The cumulative distribution of the largest eigenvalue of real Wishart random matrices can be expressed explicitly in terms of the hypergeometric function of a matrix argument. These functionals can also be described in terms of zonal polynomials. The smallest and largest eigenvalue distributions can also be expressed in terms of Tricomi functions [26]. For a detailed discussion on these objects we refer the reader to the book of Muirhead [62]. As shown in [45], these hypergeometric functions depends on alternating series involving zonal polynomials which converge very slowly even in low dimensions. Some explicit calculations for $r = 1, 2, 3$ can be found in [88].

Non-necessarily isotropic Wishart models can be considered if we restrict our attention to linear transforms and other trace-type mathematical objects. We refer to the articles of Letac and his co-authors [29, 47, 48, 91] and the tutorial [49]. See also [35, 50] for a description of the joint distribution of traces of Wishart matrices.

To bypass the complexity of finding computable and tractable closed form solutions, one natural and common method for obtaining useful information is to derive limiting distributions as the dimension tends to ∞. In this context, one can analyze the convergence of the histogram of the eigenvalues when the dimension tends to ∞. This approach is central in random matrix theory. We refer the reader to the pioneering article by E. Wigner [98] published in 1955, the lectures notes of A. Guionnet [30], the research monographs by M.L. Mehta [59] and T. Tao [89], and the references therein. This commonly used limiting theory has some drawbacks. Firstly, as the name suggests, these limiting techniques cannot capture nor control the non-asymptotic fluctuations arising in practical problems. Moreover, the limiting techniques developed in the literature often yield information only on the limiting behaviour of trace or spectral-type properties of random matrix powers. In addition, in the context of Wishart matrices, these limiting spectral-type techniques only apply to asymptotically isotropic-type models. To be more precise, the convergence analysis relies on strong hypotheses on the bias and the variance of the random matrix entries which are satisfied only for Wishart matrices with a covariance matrix close to the identity (up to some ad-hoc scaling factor). When $P \neq I$, the distribution of the eigenvalues and the corresponding eigenvectors is much more involved. The distribution of the sample eigenvalues depends on sophisticated Harish-Chandra integrals [31].

Importantly, all the spectral and trace-type approaches discussed above (whether in the limit or not) give no information on the random matrix moments themselves, but rather on their eigenvalues or trace, etc. Conversely, in many practical situations, such as in data assimilation theory and signal processing (e.g. ensemble Kalman filter theory [11, 23, 9] and particle filtering [21]), we are typically interested in the direct analysis of full matrix moments of interacting-type (non-independent) sample covariance matrices. This study concerns a step in this latter direction. Specifically, we derive formulae for the full matrix moments of real valued Wishart random matrices. As a corollary, we derive and extend a number of spectral and trace-type results for the case of non-isotropic Wishart random matrices. Laplace matrix transforms and matrix moment estimates are also studied, along with new spectral and trace concentration-type inequalities.

1.1 Organisation

Section 1.2 is concerned with the description of real random Wishart matrices and their fluctuation analysis. We also review some central result in random matrix theory, such as the semi-circle law and the Marchenko–Pastur law for isotropic Wishart matrices.

Section 1.3 provides a brief description of the main results of these notes. We provide a closed form Taylor-type formula to compute the matrix moments of P_N, and its fluctuations defined in (1.2), w.r.t. the precision parameter $1/N$. We also present the full matrix version of the semi-circle law and the Marchenko–Pastur law for non-necessarily isotropic Wishart matrices. Non-asymptotic matrix moments and exponential Laplace transforms are also provided. The last part of the section is concerned with exponential concentration inequalities for operator norms of fluctuation matrices and the eigenvalues of sample covariance matrices.

The rest of these lecture notes is concerned with the precise statement and proof of the theorems in Section 1.3. Some auxiliary outcomes and discussion surrounding these results are also given.

Section 2 reviews some useful mathematical background on Laplace and exponential inequalities, matrix norms, spectral analysis, tensor products, Fréchet derivatives, and fluctuation-type results. This section also contains a brief review of non-crossing partitions, Catalan, Narayana, and Riordan numbers, Bell polynomials, and Murasaki and circular-type representations of non-crossing partitions. The last part of this section discusses Fréchet differentiable functionals on matrix spaces and Taylor-type approximations.

Section 3 concerns closed form polynomial formulae for computing the matrix moments of P_N, and its fluctuations in (1.2), in terms of the precision parameter $1/N$ and some partition-type matrix moments. The isotropic semi-circle law (1.8) and the Marchenko–Pastur law (1.10) are simple consequences of these matrix moments expansions. New matrix versions of the semi-circle and the Marchenko–Pastur law for non-necessarily isotropic Wishart matrices are discussed in Section 3.2 and in Section 3.3.

Section 4 is concerned with some matrix moment estimates and Laplace matrix transforms of the fluctuation matrix. Spectral and trace-type concentration-type inequalities are discussed in Section 5. Section 6 is dedicated to the proof of the main theorems.

An appendix is also given containing the proof of a number of technical results required throughout these notes.

1.2 Description of the models

We recall the multivariate central limit theorem

$$P_N = P + \frac{1}{\sqrt{N}}\,\mathcal{H}_N$$

with

$$\mathcal{H}_N \overset{law}{:=} \frac{1}{\sqrt{N}} \sum_{1\leq i\leq N} (\mathbb{X}_i - P) \hookrightarrow_{N\to\infty} \mathcal{H} \tag{1.2}$$

where $\mathcal{H}$ is a symmetric $(r \times r)$-matrix with centered Gaussian entries equipped with a symmetric Kronecker covariance structure

$$(\mathcal{H}\otimes\mathcal{H})^{\sharp} = 2\,(P \,\widehat{\otimes}\, P) = \mathbb{E}\left[(\mathcal{H}_N\otimes\mathcal{H}_N)^{\sharp}\right] \tag{1.3}$$

where $(A\otimes B)^{\sharp}$ and $(A \,\widehat{\otimes}\, B)$ are the entry-wise and the symmetric tensor product of matrices A and B. These products are defined at the beginning of Section 2.3.

A detailed discussion on the fluctuation result (1.2) can be found in [41]; see also [14] for non-necessarily Gaussian variables. The fluctuation result (1.2) can also be deduced from the Laplace matrix transform estimates stated in theorem 1.5 and corollary 4.5.

Combining a perturbation analysis with the continuous mapping theorem, the central limit result (1.2) can be used to analyze the fluctuation of smooth matrix functionals of the sample covariance matrix.

Roughly speaking, given some smooth Fréchet differentiable mapping $\Upsilon : \mathcal{S}_r \mapsto \mathcal{B}$ from symmetric matrices $\mathcal{S}_r$ to some Banach space $\mathcal{B}$, we have the Taylor expansion

$$\begin{aligned}\mathcal{H}_N^{\Upsilon} &:= \sqrt{N}\,\left[\Upsilon(P_N) - \Upsilon(P)\right] \\ &= \nabla\Upsilon(P)\cdot\mathcal{H}_N + \frac{1}{2\sqrt{N}}\,\nabla^2\Upsilon(P)\cdot(\mathcal{H}_N,\mathcal{H}_N) + \ldots\end{aligned}$$

Using the unbiasedness properties of the sample covariance matrix, the second order term gives the bias of the estimate $\Upsilon(P_N)$; that is we have that

$$\mathbb{E}\left[\mathcal{H}_N^{\Upsilon}\right] = \frac{1}{2\sqrt{N}}\,\mathbb{E}\left[\nabla^2\Upsilon(P)\cdot(\mathcal{H},\mathcal{H})\right] + \mathrm{O}\left(\frac{1}{N}\right)$$

Equivalently, we have

$$\mathbb{E}\left[\Upsilon(P_N)\right] = \Upsilon(P) + \frac{1}{2N}\,\mathbb{E}\left[\nabla^2\Upsilon(P)\cdot(\mathcal{H},\mathcal{H})\right] + \mathrm{O}\left(\frac{1}{N^{3/2}}\right)$$

For a more precise statement and several illustrations we refer the reader to Section 2.6, theorem 2.2. For instance, for power functions $\Upsilon_n(Q) := Q^n$, for any $1 \leq m \leq n$ we have

$$\begin{aligned}&\mathbb{E}\left[\nabla^m\Upsilon_n(P)\cdot\mathcal{H}_N^{\otimes m}\right] \\ &= m!\sum_{0\leq i_1<\ldots<i_m\leq n}\mathbb{E}\left(\prod_{1\leq k\leq m}\left[P^{i_k-i_{k-1}-1}\,\mathcal{H}_N\right]\right) \\ &= (n)_m\,\mathbb{E}(\mathcal{H}_N^m)\quad\text{when}\quad P = I\end{aligned}\tag{1.4}$$

with the Pochhammer symbol $(n)_m := n!/(n-m)!$, and the convention $(i_0, i_{m+1}) = (0, n)$. In this context, the m-moments of the fluctuation matrices $\mathcal{H}_N$ represents the mean-error of order m. This property also holds for rational powers. For instance, we have the non-asymptotic estimate on the Frobenius norm,

$$\left\| \mathbb{E}\left(\sqrt{P_N}\right) - \sqrt{P} + \frac{1}{2N}\,\sqrt{P}\,\left[\frac{1}{4}\,I + \int_0^\infty t\,\mathrm{Tr}\left(P\,e^{-t\sqrt{P}}\right)\,e^{-t\sqrt{P}}\,dt\right]\right\|_F \leq \frac{r}{4}\,\frac{1}{N\sqrt{N}}\,\lambda_{min}(P)^{-5/2}\,\left[\mathrm{Tr}(P^2) + \mathrm{Tr}(P)^2\right] \tag{1.5}$$

The proof of this assertion is provided in Section 2.6.

To summarise the consequences of the preceding discussion, to analyze these approximations at any order it is therefore necessary to be able to compute the m-moments of the fluctuation matrices $\mathcal{H}_N$.

Gaussian approximation techniques also require one to estimate the fluctuations of the moment $\mathbb{E}(\mathcal{H}_N^m)$ around those of $\mathbb{E}(\mathcal{H}^m)$ given in terms of the limiting Gaussian matrix $\mathcal{H}$, and with respect to the sample size parameter. Moreover, one often wants to control the behavior of these objects when the dimension parameter tends to ∞.

The limiting random matrix model $\mathcal{H}$ discussed above is closely related to Gaussian orthogonal ensembles arising in random matrix theory. To be more precise, we can check that

$$\mathcal{H} \stackrel{law}{=} P^{1/2}\left(\frac{\mathcal{W}+\mathcal{W}'}{\sqrt{2}}\right)P^{1/2} \tag{1.6}$$

where $\mathcal{W} = (\mathcal{W}_{i,j})_{1\leq i,j\leq r}$ is a matrix of independent, centred Gaussian elements of unit variance.

When $P = I$, the random matrix $\mathcal{H}$ introduced in (1.6) reduces to a Gaussian orthogonal ensemble. In this situation, we have

$$r^{-2}\,\mathbb{E}\left(\mathrm{Tr}\left[\mathcal{H}^2\right]\right) = 1 + r^{-1} \quad \text{and} \quad r^{-3}\,\mathbb{E}\left(\mathrm{Tr}\left[\mathcal{H}^4\right]\right) = 2 + 5\,r^{-2} + 5\,r^{-1} \tag{1.7}$$

The trace of the higher moments $\mathbb{E}\left(\mathcal{H}^n\right)$ can be estimated using the semi-circle law (1.8) in large dimensions. This celebrated limiting result is proved using the convergence of moment property

$$r^{-1}\,\mathbb{E}\left(\mathrm{Tr}\left(\left[\frac{\mathcal{H}}{\sqrt{r}}\right]^n\right)\right) = 1_{2\mathbb{N}}(n)\; C_{n/2} + O\left(1/r\right) \qquad (1.8)$$
$$= \frac{1}{2\pi}\int_{-2}^{2} x^n\,\sqrt{4-x^2}\;dx + O\left(1/r\right)$$

for any $n \geq 1$, with the Catalan numbers

$$C_n := \frac{1}{n+1}\begin{pmatrix} 2n \\ n \end{pmatrix} \qquad (1.9)$$

A proof of the above assertion via Wick's theorem, including detailed reference pointers is given in [30, Section 1.4], see also [60, Chapter 1].

When $P = I$ and $N = r/\rho$ for some parameter $\rho > 0$, another important result is the Marchenko–Pastur law

$$\lim_{r\to\infty} r^{-1}\,\mathbb{E}\left[\mathrm{Tr}\left(P_N^n\right)\right]$$
$$= \sum_{0\leq m<n} \frac{\rho^m}{m+1}\begin{pmatrix} n \\ m \end{pmatrix}\begin{pmatrix} n-1 \\ m \end{pmatrix}$$
$$= \int_{a_-(\rho)}^{a_+(\rho)} x^n\left[\left(1-\frac{1}{\rho}\right)_+ \delta_0(dx) + \frac{1}{2\pi\rho\, x}\sqrt{\left[a_+(\rho)-x\right]\left[x-a_-(\rho)\right]}\;dx\right]$$

with the parameters

$$a_-(\rho) := (1-\sqrt{\rho})^2 \quad \text{and} \quad a_+(\rho) := (1+\sqrt{\rho})^2$$

The proof of the above integral formula can be found in [101, lemma 5.2], see also [28, 65] and the pioneering article by Vladimir Marchenko and Leonid Pastur [56]. A new proof of this result follows from the full matrix version of the Marchenko–Pastur law given in corollary 3.3 of these notes.

When $P \neq I$, formula (1.6) can be combined with Isserlis' theorem [33] (or Wick's theorem [97]) to compute the matrix moments of the random matrix $\mathcal{H}$. For instance, we have $\mathbb{E}(\mathcal{H}^{2n+1}) = 0$, for any $n \geq 0$. After some lengthy combinatorial computations we also find the matrix polynomials

$$\begin{aligned}\mathbb{E}\left(\mathcal{H}^2\right) &= P^2 + \mathrm{Tr}(P)\,P \\ \mathbb{E}\left(\mathcal{H}^4\right) &= 5\,P^4 + 3\,\mathrm{Tr}(P)\,P^3 + \left[\mathrm{Tr}(P^2) + \mathrm{Tr}(P)^2\right]\,P^2 \\ &\qquad + \left[\mathrm{Tr}(P^3) + \mathrm{Tr}(P)\,\mathrm{Tr}(P^2)\right]\,P\end{aligned}$$

Although these matrix moments are given by a closed form formula, their complex combinatorial structure cannot be used in simple calculations. For example, the calculation of $\mathbb{E}\left(\mathcal{H}^{2n}\right)$ requires the matrix moments associated with $2^{-n}\,(2n)!/n!$ partitions over $[2n] := \{1,\ldots,2n\}$ with n-blocks. For $n = 4$, more than one hundred moments need to be computed. The computational complexity to numerically compute the central moments of the multivariate normal distribution is discussed in [68]; see also [3, p. 49], [38, proposition 1], [62, p. 46], and the matrix derivative formula in [92].

The above formulae also show that we cannot expect to have a semi-circle-type law as in (1.7) for any covariance matrix. Different types of behaviour can be expected depending on the behavior of the eigenvalues of P w.r.t. the dimension parameter r. For instance, if the largest eigenvalue is $\lambda_1(P) = r$, we have $1 \le r^{-1}\,\mathrm{Tr}(P) \le 2$ but

$$\mathbb{E}\left(\mathrm{Tr}\left[\mathcal{H}^4\right]\right) \ge 5\,r^4 \implies r^{-1}\,\mathbb{E}\left(\mathrm{Tr}\left(\left[\frac{\mathcal{H}}{\sqrt{r}}\right]^4\right)\right) \longrightarrow_{r\to\infty} \infty$$

1.3 Statement of some main results

One of the main objectives of these lecture notes is to analyze the properties of real Wishart matrix moments. Let $\mathcal{P}_n$ be the set of all partitions π of $[n] := \{1,\ldots,n\}$, $\mathcal{P}_{n,m} \subset \mathcal{P}_n$ be the subset of all partitions with m blocks $\pi_1 \le \ldots \le \pi_m$ ordered in a canonical way w.r.t. their smallest element.

Let $\mathcal{Q}_n \subset \mathcal{P}_n$ and $\mathcal{Q}_{n,m} \subset \mathcal{P}_{n,m}$ be the subset of partitions without the singleton. Also let $\alpha^\pi := \sum_{1\le i\le m} i\ 1_{\pi_i}$. In other words, $\alpha^\pi(i)$ is the index of the block of π containing index i.

The π-matrix moments $M_\pi^{[Q]}(P)$ and $M_\pi^{\circ,[Q]}(P)$ associated with some collection of $(r \times r)$ matrices $(Q_i)_{i \geq 1}$ are defined by

$$M_\pi^{[Q]}(P) := \mathbb{E}\left([\mathbb{X} - P]_\pi^Q\right) \quad \text{and} \quad M_\pi^{\circ,[Q]}(P) := \mathbb{E}\left(\mathbb{X}_\pi^Q\right) \tag{1.10}$$

with the random matrices

$$\mathbb{X}_\pi^Q := \prod_{1 \leq i \leq n} \left[Q_i \mathbb{X}_{\alpha^\pi(i)}\right]$$

and

$$[\mathbb{X} - P]_\pi^Q := \prod_{1 \leq i \leq n} \left[Q_i (\mathbb{X}_{\alpha^\pi(i)} - P)\right]$$

We also consider the matrix moments

$$M_{n,m}^{[Q]}(P) := \sum_{\pi \in \mathcal{Q}_{n,m}} M_\pi^{[Q]}(P) \quad \text{and} \quad M_{n,m}^{\circ,[Q]}(P) := \sum_{\pi \in \mathcal{P}_{n,m}} M_\pi^{\circ,[Q]}(P)$$

Our first main result provides polynomial formulae w.r.t. the precision parameter $1/N$.

Theorem 1.1. For any collection of matrices Q_n, and any $2N \geq n \geq 1$, we have the polynomial formulae

$$\mathbb{E}\left[(Q_1 \mathcal{H}_N) \ldots (Q_n \mathcal{H}_N)\right] = \sum_{1 \leq m \leq \lfloor n/2 \rfloor} \frac{1}{N^{n/2-m}} \, \partial_{n,m}^{[Q]}(P) \tag{1.11}$$

with

$$\partial_{n,m}^{[Q]}(P) := \sum_{m \leq l \leq \lfloor n/2 \rfloor} s(l,m) \; M_{n,l}^{[Q]}(P)$$

In addition, we have

$$\mathbb{E}\left[(Q_1 P_N) \ldots (Q_n P_N)\right] = \sum_{1 \leq m \leq n} \frac{1}{N^{n-m}} \, \partial_{n,m}^{\circ,[Q]}(P) \tag{1.12}$$

with

$$\partial_{n,m}^{\circ,[Q]}(P) = \sum_{m \leq l \leq n} s(l,m) \; M_{n,l}^{\circ,[Q]}(P)$$

In the above displayed formulae, $s(l,m)$ are the Stirling numbers of the first kind.

For the detailed discussion of these matrix moments, including several corollaries and examples, we refer to Section 3.1; see for example theorem 3.1 when $Q_i = I$.

To simplify notation, for homogeneous models $Q_i = I$ and sequences of matrices $P : r \mapsto P(r)$ we suppress the indices $(.)^{[I]}$ and r, and write

$$(\partial_{n,m}(P), \partial^{\circ}_{n,m}(P), M_{\pi}(P), M^{\circ}_{\pi}(P), M_{n,l}(P), M^{\circ}_{n,l}(P))$$

instead of

$$(\partial^{[I]}_{n,m}(P(r)), \partial^{\circ,[I]}_{n,m}(P(r)), M^{[I]}_{\pi}(P), M^{\circ,[I]}_{\pi}(P), M^{[I]}_{n,l}(P(r)), M^{\circ,[I]}_{n,l}(P(r)))$$

The polynomial formula (1.11) differs from the invariant moments which can be derived using the algorithm presented in [47]. In the latter, the authors provide an elegant spectral technique to interpret these moments in terms of spherical polynomials and matrix-eigenfunctions of Wishart integral operators; see [47, e.g. proposition 4.3]. A drawback of this spectral method is that it requires one to diagonalize and invert complex combinatorial matrices. It is difficult to use this technique to derive estimates w.r.t. the sample size parameter. Matrix moment formulae can also be derived from [73]. Nevertheless the resulting Isserlis-type decompositions will involve complex series of summations over pair partitions.

Beside the fact that the matrix moments $M^{[Q]}_{n,l}(P)$ can be computed using Isserlis' theorem, to be the best of our knowledge no explicit and closed form polynomial formulae in terms of P are known. In the further development of these notes, we provide estimates of the fluctuation matrix moments w.r.t. the sample size in terms of the dominating term of the sum (1.11). A brief description of these estimates are provided in theorem 1.5 below.

To move one step further in our discussion we assume that $Q_i = I$ and $n = 2m$. In this situation the single dominating term in (1.11) is given by the central matrix moments

$$\partial_{2m,m}(P) = M_{2m,m}(P) \tag{1.13}$$

This implies that

$$\mathbb{E}\left[\mathcal{H}^{2m}_N\right] = M_{2m,m}(P) + \mathrm{O}\left(\frac{1}{N}\right) I$$

We say a partition is a crossing partition whenever we can find $i < j < k < l$ with i, k in a block and j, l in the other block. Let $\mathcal{N}_n \subset \mathcal{P}_n$ and $\mathcal{N}_{n,m} \subset \mathcal{P}_{n,m}$ be the subsets of non-crossing partitions.

We denote by $\Sigma_n(P)$ the matrix polynomial given by

$$\Sigma_n(P) \ := \ \mathrm{Tr}(P) \sum_{\pi \in \mathcal{N}_n} \left[\prod_{i \geq 0} \mathrm{Tr}\left(P^{1+i}\right)^{r_i(\pi)} \right] \frac{P^{|\pi_1|}}{\mathrm{Tr}(P^{|\pi_1|+1})} \tag{1.14}$$

In the above display, $r_0(\pi) := n - \sum_{i \geq 1} r_i(\pi)$ where $r_i(\pi)$ is the number of blocks of size $i \geq 1$ in the partition π.

Also let $\Sigma^{\circ}_{n,m}(P)$ be the matrix polynomial given by

$$\Sigma^{\circ}_{n,m}(P) \ := \ \sum_{\pi \in \mathcal{N}_{n,m}} \left[\prod_{i \geq 1} \mathrm{Tr}(P^i)^{r_i(\Xi(\pi))} \right] \frac{P^{\iota(\pi)}}{\mathrm{Tr}\left(P^{\iota(\pi)}\right)} \tag{1.15}$$

In the above display, $\iota(\pi)$ denotes the number of blocks visible from above in the Murasaki diagram associated with π; see Section 2.5 for examples. The partition $\Xi(\pi) \in \mathcal{N}_{n+1-m}$ is defined in terms of a circular representation of π. That is, firstly, we subdivide the n arcs of $\pi \in \mathcal{N}_{n,m}$ by a new series of n nodes placed clockwise. Then $\Xi(\pi)$ is the coarsest non-crossing partition of these nodes whose chords don't cross those of π. For a detailed description of the mapping Ξ, and examples, we refer Section 2.5; see e.g. (2.20).

Lets further assume that $P : r \mapsto P(r)$ is a collection of possibly random matrices satisfying for any $n \geq 1$ the almost sure convergence of the moments

$$r^{-1} \ \tau_n \left(P(r)\right) := r^{-1} \ \mathrm{Tr}(P(r)^n) \quad \longrightarrow_{r \to \infty} \quad \tau_n(P) \tag{1.16}$$

Also, let $\mathcal{H}_N(r)$ and $\mathcal{H}(r)$ be the random matrix model defined as in (1.2) and (1.6) by replacing P by $P(r)$.

To simplify notation, we write

$$(\mathcal{H}, \mathcal{H}_N, M_{2n,n}(P), M^{\circ}_{n,m}(P), \Sigma_n(P), \Sigma^{\circ}_{n,m}(P))$$

instead of

$$(\mathcal{H}(r), \mathcal{H}_N(r), M_{2n,n}(P(r)), M^{\circ}_{n,m}(P(r)), \Sigma_n(P(r)), \Sigma^{\circ}_{n,m}(P(r)))$$

In this notation, the next theorem relates the matrix moments $(\Sigma_n(P), \Sigma^\circ_{n,m}(P))$ with the matrix moments $(M_{n,m}(P), M^\circ_{n,m}(P))$ and the ones of the Gaussian matrix $\mathcal{H}$.

Theorem 1.2. Let $P : r \mapsto P(r)$ be a collection of possibly random matrices satisfying the condition (1.16). In this situation, the central matrix moments $M_{2n,n}(P)$ coincide with the ones of the limiting Gaussian matrix. In addition, for any $n \geq m \geq 1$ we have the matrix moment estimates

$$\begin{aligned} M_{2n,n}(P) =& \ \mathbb{E}\left(\mathcal{H}^{2n}\right) \ = \ \Sigma_n(P) + \mathrm{O}\left(r^{n-1}\right) I \\ M^\circ_{n,m}(P) =& \ \Sigma^\circ_{n,m}(P) + \mathrm{O}\left(r^{n-m-1}\right) I \end{aligned} \tag{1.17}$$

For a proof and a more detailed discussion on these matrix moment relations we refer to Section 3.2 and Section 3.3; see e.g. theorem 3.4, theorem 3.5 and theorem 3.7.

The first line estimate in (1.17) is a consequence of the decomposition (3.1) and theorem 3.4. The second line estimate in (1.17) is a consequence of the estimates (3.18) and theorem 3.7.

Theorem 1.2 together with (1.13) yields the estimates

$$r^{-(n+1)}\ \mathbb{E}\left(\mathcal{H}^{2n}\right) = \overline{\Sigma}_n(P) + \mathrm{O}\left(r^{-1}\right) I_r$$

as well as

$$\mathbb{E}\left[\mathcal{H}_N^{2n}\right] = \mathbb{E}\left(\mathcal{H}^{2n}\right) + \mathrm{O}\left(N^{-1}\right) I$$

We also have

$$r^{-(n-m+1)} M^\circ_{n,m}(P) = \overline{\Sigma}^\circ_{n,m}(P) + \mathrm{O}\left(r^{-1}\right) I_r \quad \text{with} \quad I_r := r^{-1}\, I$$

with the matrix polynomials $(\overline{\Sigma}_n(P), \overline{\Sigma}^\circ_{n,m}(P))$ defined similarly to $\left(\Sigma_n(P), \Sigma^\circ_{n,m}(P)\right)$ but with the trace operator replaced by the normalized traces

$$\overline{T}\mathrm{r}(Q) := r^{-1}\, \mathrm{Tr}(Q)$$

A more refined estimate between $\mathbb{E}\left[\mathcal{H}_N^{2n}\right]$ and $\mathbb{E}\left(\mathcal{H}^{2n}\right)$ can be found later in theorem 1.5.

The first line assertion in (1.17) in theorem 1.2 provides a semicircle-type asymptotic theorem when the dimension parameter tends to ∞.

Corollary 1.3. Under the assumptions of theorem 1.2, we have the extended semi-circle law

$$r^{-1}\ \mathbb{E}\left(\mathrm{Tr}\left(\left[\frac{\mathcal{H}}{\sqrt{r}}\right]^{2n}\right)\right)\ = \sigma_n(P)\ + O\left(r^{-1}\right)$$

with

$$\sigma_n(P) := 2\sum_{\mu}\binom{n}{\mu_1\ \mu_2\ \ldots\ \mu_n}\ \tau_\mu(P) \quad \text{and} \quad \tau_\mu(P) := \prod_{1\leq i\leq n}\tau_i(P)^{\mu_i} \tag{1.18}$$

In the above display, the summation is taken over all collection of non-negative indices $\mu = (\mu_1, \ldots, \mu_n)$ such that

$$\sum_{1\leq i\leq n}\mu_i = n+1 \quad \text{and} \quad \sum_{1\leq i\leq n} i\mu_i = 2n \tag{1.19}$$

For instance, we have

$$\begin{aligned}
\sigma_1(P) &= \tau_1(P)^2 \\
\sigma_2(P) &= 2\ \tau_1(P)^2\ \tau_2(P) \\
\sigma_3(P) &= 2\ \tau_1(P)^3\ \tau_3(P) + 3\ \tau_1(P)^2\ \tau_2(P)^2 \\
\sigma_4(P) &= 2\ \tau_1(P)^4\ \tau_4(P) + 8\ \tau_1(P)^3\ \tau_2(P)\ \tau_3(P) + 4\ \tau_1(P)^2\ \tau_2(P)^3 \\
\sigma_5(P) &= 2\ \tau_1(P)^5\ \tau_5(P) + 10\ \tau_1(P)^4\ \tau_2(P)\ \tau_4(P) + 5\ \tau_1(P)^4\ \tau_3(P)^2 \\
&\quad +\ 20\ \tau_1(P)^3\ \tau_2(P)^2\ \tau_3(P) + 5\ \tau_1(P)^2\ \tau_2(P)^4
\end{aligned}$$

Observe that $\sigma_n(\alpha\, I) = C_n\ \alpha^{2n}$, for any $\alpha \geq 0$. These formulae can be checked combining (3.16) and (3.17) with corollary 3.6. Matrix-valued free probability techniques can also be used to recover the above matrix moment formula [66, 86, 87, 90, 95].

To the best of our knowledge the extended and matrix version of the semi-circle law stated in the above theorem have not been considered in the literature. See also Section 3.2.

Also recall that Carleman's condition

$$\sum_{n\geq 1}\sigma_n(P)^{-1/(2n)} = \infty$$

ensures the existence and uniqueness of a random variable with null odd moments and the $(2n)$-moments $\sigma_n(P)$ defined in corollary 1.3

(cf. [1, 16] and p. 296 in [78]). For instance, when $P = I$ we have

$$\sigma_n(I) = C_n \simeq \frac{2^{2n}}{n^{3/2}\sqrt{\pi}} \leq 2^{2n}$$

This implies that

$$\sigma_n(I)^{-1/(2n)} \geq \frac{1}{2} \quad \Longrightarrow \quad \sum_{n\geq 1} \sigma_n(I)^{-1/(2n)} = \infty$$

In this case, the random variable with null odd moments and the $(2n)$-moments $\sigma_n(I)$ is given by the semi-circle law (1.8).

Another direct consequence of theorem 1.2 is the Marchenko–Pastur law for non-isotropic Wishart matrices due to Y.Q. Yin [102, 103]; see also [18] and [80].

Corollary 1.4 ([102, 103]). Consider a collection $P : r \mapsto P(r)$ of possibly random matrices satisfying the condition stated in (1.16). Let $N = r/\rho$ be a scaling of the sample size in terms of the dimension associated with some parameter $\rho > 0$. For any $n \geq 1$ we have the Kreweras-type formula

$$\lim_{r\to\infty} r^{-1} \operatorname{Tr}\left(\mathbb{E}\left[P_N^n\right]\right) = \sum_{1\leq m\leq n} \rho^{n-m} \sum_{\mu\vdash[n]\ :\ m+|\mu|=n+1} K(\mu)\ \tau_\mu(P)$$

with the trace parameters $\tau_\mu(P)$ and the Kreweras numbers $K(\mu)$ defined in (1.18) and later in (2.13).

When $P = I$ the above limit result reduces to

$$\lim_{r\to\infty} r^{-1}\ \mathbb{E}\left[\operatorname{Tr}\left(P_N^n\right)\right] = \sum_{1\leq m\leq n} \rho^{n-m}\ N_{n,m}$$

In this situation, we also have the centered version limiting result

$$\lim_{r\to\infty} r^{-1}\mathbb{E}\left(\operatorname{Tr}\left([P_N - I]^n\right)\right) = \sum_{1\leq m\leq \lfloor n/2\rfloor} \rho^{n-m}\ R_{n,m}$$

In the above display, $N_{n,m}$ and $R_{n,m}$ denote the Narayana and the Riordan numbers defined later in (2.13) and (2.15). The matrix version of these isotropic results can be found in corollary 3.3 with the trace-type Marchenko–Pastur law (1.10) a simple corollary. See Section 3.3 for a new matrix version of a non-isotropic Marchenko–Pastur law.

Our third main result concerns moment estimates. We let $\|.\|_{op}$ and $\|.\|_F$ denote the operator norm and the Frobenius norm. In this notation, we have the following theorem.

Theorem 1.5. For some sufficiently small time horizon and for any sufficiently large sample size and any $n \geq 1$ we have the estimates

$$\begin{aligned} N \, \|\mathbb{E}\left[\mathcal{H}_N^{2n}\right] - \mathbb{E}\left[\mathcal{H}^{2n}\right] \|_F &\leq c_1^n \; (2n)_n \; \mathrm{Tr}(P)^{2n} \\ \sqrt{N} \, \|\mathbb{E}\left(\exp\left(t\mathcal{H}_N\right)\right) - \mathbb{E}\left(\exp\left(t\mathcal{H}\right)\right) \|_F &\leq c_2 \; (t \, \mathrm{Tr}(P))^3 \\ \mathbb{E}\left[\|\mathcal{H}\|_{op}\right] \wedge \mathbb{E}\left[\|\mathcal{H}_N\|_{op}\right] &\leq c_3 \; \sqrt{r} \; \lambda_1(P) \end{aligned}$$

for some finite universal constants $c_1, c_2, c_3 < \infty$ whose values do not depend on the dimension parameter, nor on the parameter n.

In the above display, $\lambda_1(P) = \|P\|_{op}$ denotes the maximal eigenvalue of P (cf. 2.2).

A more precise statement with a more detailed description of the constants is provided in Section 4 and Section 5; see for instance theorem 4.1, theorem 4.4, corollary 4.5, and theorem 5.4. The operator norm estimate stated in the above theorem extends the norm estimate for isotropic random vectors presented in [75] in the context of Gaussian random matrices. These norm-type bounds are based on non-commutative versions of Khintchine-type inequalities for Rademacher series presented in [54, 55]. More sophisticated approaches based on Burkholder/Rosenthal martingale-type inequalities are also developed in [36, 37]. Nevertheless these inequalities cannot be used to estimate random operator norms and the constants are often not explicit.

The last part of these lecture notes is concerned with non-asymptotic exponential concentration inequalities for traces and the operator norm of the fluctuation matrix. In this context, our main results can be stated as follows.

Theorem 1.6. For any symmetric matrix A, any $\delta \geq 0$, and any sufficiently large sample size the probabilities of the following events

$$\begin{aligned} |\mathrm{Tr}(A\mathcal{H}_N)| &\leq c_1 \sqrt{(\delta+1) \;\; [\mathrm{Tr}((AP)^2) + \|AP\|_F^2]} \\ \|\mathcal{H}_N\|_{op} &\leq \;\; c_2 \; \lambda_1(P) \; \sqrt{\delta + r} \\ \sup_{1\leq k\leq r} |\lambda_k(P_N) - \lambda_k(P)| &\leq \;\; c_2 \; \lambda_1(P) \; \sqrt{(\delta+r)/N} \end{aligned}$$

are greater than $1 - e^{-\delta}$, where c_1, c_2 denote some universal constants.

In the above display, $\lambda_i(P_N)$ and $\lambda_i(P)$ denote the ordered (decreasing in magnitude) eigenvalues of P_N, resp. P (cf. 2.2). For a precise statement of this result and a detailed description of the constants c_1, c_2 we refer to Section 5; see in particular Section 5.2 and theorem 5.2, theorem 5.4 and corollary 5.5.

The proof of the trace-type concentration inequality is based on sub-Gaussian Laplace estimates of well known Wishart trace-type Laplace transforms (5.5). See Section 5.2 for a description of these sub-Gaussian estimates; e.g. see (5.7) and the first assertion in theorem 5.2. The operator norm concentration inequality comes from the variational formulation

$$\|\mathcal{H}_N\|_{op} = \sup_{x,y \in \mathbb{B}} \langle \mathcal{H}_N x, y \rangle = \sup_{A \in \mathbb{A}} \mathrm{Tr}(A\mathcal{H}_N) \tag{1.20}$$

where $\mathbb{B}$ is the unit ball in $\mathbb{R}^r$ equipped with the Euclidian distance and $\mathbb{A}$ is the set of matrices

$$\mathbb{A} := \{A = (xy' + y'x)/2 \ : \ x, y \in \mathbb{B}\} \tag{1.21}$$

The last spectral concentration estimate is a direct consequence of Weyl's inequality (2.5).

We end this section with some comparisons of the above concentration inequalities with existing results in random matrix theory. When $P = I$ the joint density of the random eigenvalues of P_N is explicitly known; see e.g. [4]. Elegant Sanov-type large deviation principles for the spectral empirical measures have been developed by G. Ben Arous and A. Guionnet [5]. The literature also consists of non-asymptotic concentration inequalities for sums of independent random matrices. We refer to the seminal book of J. Tropp [93] for the state of the art on these topics. See also the review [94].

We also emphasize that the Laplace transform-type techniques developed in the present study differ from the ones based on Lieb's inequality (4.13). The latter are often used to control the largest eigenvalue of a random matrix using trace-type estimates; see proposition 4.4 and Section 4.5 in [93].

Other types of models have been considered in the literature leading to different results. For example, Gaussian concentration inequalities have been derived for Rademacher and Gaussian series associated with deterministic self-adjoint matrices; see e.g. theorem 2.1 in [93]. Matrix Hoeffding, Bernstein and Azuma-type inequalities have been derived for almost surely bounded random matrices; see theorem 2.8 and theorem 8.1 in [93]. The concentration results developed in the present notes provide more refined estimates, but of course they are restricted to random Wishart matrix models.

References

[1] N.I. Akhiezer. The Classical Moment Problem and Some Related Questions in Analysis. Oliver & Boyd (1965).

[2] F.P. Anaraki and S. Hughes. Memory and computation efficient PCA via very sparse random projections. In International Conference on Machine Learning, pp. 1341–1349 (2014).

[3] T.W. Anderson. An Introduction to Multivariate Statistical Analysis. 3rd Edition. John Wiley & Sons, New York (2003).

[4] G.W. Anderson, A. Guionnet, and O. Zeitouni. An Introduction to Random Matrices. Cambridge University Press (2010).

[5] G. Ben Arous and A. Guionnet. Large deviations for Wigner's law and Voiculescu's non-commutative entropy. Probability Theory and Related Fields. vol. 108, no. 4. pp. 517–542 (1997).

[6] F.R. Bernhart. Catalan, Motzkin, and Riordan numbers. Discrete Mathematics. vol. 204, no. 1–3. pp. 73–112 (1999).

[7] E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 245–250. ACM (2001).

[8] C.M. Bishop. Pattern Recognition and Machine Learning. Springer (2006).

[9] A.N. Bishop and P. Del Moral. Matrix product moments in normal variables. arXiv e-print, arXiv:1808.00235 (2018).

[10] A.N. Bishop and P. Del Moral. On the Stability of Matrix-Valued Riccati Diffusions. arXiv e-print, arXiv:1703.00353 (2017).

[11] A.N. Bishop, P. Del Moral, and A. Niclas. A perturbation analysis of stochastic matrix Riccati diffusions. arXiv e-print, arXiv:1709.05071 (2017).

[12] M. Bóna and I. Mezö. Real zeros and partitions without singleton blocks. European Journal of Combinatorics. vol. 51. pp. 500–510 (2016).

[13] J.P. Bouchaud and M. Potters. Theory of Financial Risk and Derivative Pricing From Statistical Physics to Risk Management. Cambridge University Press (2009).

[14] M.W. Browne. Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology. vol. 37, no. 1. pp. 62–83 (1984).

[15] M.F. Bru. Wishart Processes. Journal of Theoretical Probability. vol. 4, no. 4. pp. 725–751 (1991).

[16] T. Carleman. Les fonctions quasi analytiques. Collection Borel, Gauthier–Villars, Paris (1926).

[17] X. Chen and W. Chu. Moments on Catalan numbers. Journal of Mathematical Analysis and Applications. vol. 349, no. 2. pp. 311–316 (2009).

[18] S.I. Choi. The inversion formula of the Stieltjes transform of spectral distribution. Journal of the Chungcheong Mathematical Society. vol. 22, no. 3. pp. 519–524 (2009).

[19] J.C. Cox, J.E. Ingersoll, and S.A. Ross. A Theory of the Term Structure of Interest Rate. Econometrica. vol. 53, no. 2. pp. 385–407 (1985).

[20] C. Cuchiero, D. Filipovich, E. Mayerhofer, and J. Teichman. Affine processes on positive semidefinite matrices. The Annals of Applied Probability. vol. 21, no. 2. pp. 397–463 (2011)

[21] P. Del Moral. Feynman-Kac formulae: Genealogical and interacting particle systems with applications. Springer, New York (2004).

[22] P. Del Moral. Mean field simulation for Monte Carlo integration. Chapman & Hall/CRC Monographs on Statistics & Applied Probability (2013).

[23] P. Del Moral and J. Tugaut. On the stability and the uniform propagation of chaos properties of ensemble Kalman-Bucy filters. Annals of Applied Probability. vol. 28, no. 2. pp. 790–850 (2018).

[24] N. Dershowitz and S. Zaks. Ordered trees and non-crossing partitions. Discrete Mathematics. vol. 62, no. 1. pp. 215–218 (1986).

[25] T.A. Dowling. Stirling numbers. In Applications of Discrete Mathematics (eds: J.G. Michaels and K.H. Rosen). Chapter 6. McGraw-Hill (1991).

[26] A. Edleman. Eigenvalues and condition numbers of random matrices. Ph.D Thesis, MIT (1989).

[27] E.J. Elton and M.J. Gruber. Modern Portfolio Theory and Investment Analysis. John Wiley and Sons, New York (1995).

[28] F. Götze and A. Tikhomirov. Rate of convergence in probability to the Marchenko–Pastur law. Bernoulli. vol. 10, no. 3, pp. 503–548 (2004).

[29] P. Graczyk, G. Letac, and H. Massam. The hyperoctahedral group, symmetric group representations and the moments of the real Wishart distribution. Journal of Theoretical Probability. vol. 18, no. 1. pp. 1–42 (2005).

[30] A. Guionnet. Large Random Matrices: Lectures on Macroscopic Asymptotics. Lecture Notes in Mathematics. Springer–Verlag (2009).

[31] Harish-Chandra. Differential operators on a semi-simple Lie algebra. American Journal of Mathematics. vol. 79, no.1. pp. 87—120 (1957).

[32] H. Epstein. Remarks on Two Theorems of E. Lieb, Comm. Math. Phys., vol. 31, pp. 317–325 (1973).

[33] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika, vol. 12, no. 1/2. pp. 134–139 (1918).

[34] O. James, H-N. Lee. Concise Probability Distributions of Eigenvalues of Real-Valued Wishart Matrices. arXiv e-print, arXiv:1402.6757 (2014).

[35] R.D. Jensen. The joint distribution of traces of Wishart matrices and some applications. The Annals of Mathematical Statistics. vol. 41, no. 1. pp. 133–145 (1970).

[36] M. Junge and Q. Xu. On the best constants in some non-commutative martingale inequalities. Bulletin of the London Mathematical Society. vol. 37, no. 2. pp. 243–253 (2005).

[37] M. Junge and Q. Xu. Noncommutative Burkholder/Rosenthal inequalities II: Applications. Israel Journal of Mathematics. vol. 167, no 1. pp. 227–282 (2008).

[38] R. Kan. From Moments of Sums to Moments of Product. Journal of Multivariate Analysis, vol. 99, no. 3. pp. 542–554 (2008).

[39] N.D. Kazarinoff. Geometric inequalities. Random House, New York (1961).

[40] M. Keller-Ressel, W. Schachermayer, and J. Teichmann. Affine processes are regular. Probability Theory and Related Fields. vol. 151. no. 3-4. pp. 591–611 (2011).

[41] M.G. Kendall and A. Stuart. The Advanced Theory of Statistics. C. Griffin & Company, London (1943).

[42] J.S. Kim, K.H. Lee, and S.J. Ho. Weight multiplicities and Young tableaux through affine crystals. arXiv e-print, arXiv:1703.10321 (2017).

[43] I. Kortchemski and C. Marzouk. Simply generated non-crossing partitions. Combinatorics, Probability and Computing. vol. 26, no. 4. pp. 560–592 (2017).

[44] G. Kreweras. Sur les partitions non croisées d'un cycle. Discrete Mathematics. vol. 1, no. 4. pp. 333–350 (1972).

[45] P.R. Krishnaiah and T.C. Chang. On the exact distribution of the smallest root of the Wishart matrix using zonal polynomials. Annals of the Institute of Statistical Mathematics. vol. 23, no. 11. pp. 293–295 (1971).

[46] M.J. Kusner, N. I. Kolkin, S. Tyree, and K.Q. Weinberger. Image Data Compression for Covariance and Histogram Descriptors. arXiv preprint arXiv:1412.1740 (2014).

[47] G. Letac and H. Massam. All invariant moments of the Wishart distribution. Scandinavian Journal of Statistics. vol. 31, no. 2. pp. 295–318 (2004).

[48] G. Letac and H. Massam. The noncentral Wishart as an exponential family, and its moments. Journal of Multivariate Analysis. vol. 99. no. 7. pp. 1393–1417 (2008).

[49] G. Letac and H. Massam. A tutorial on non central Wishart distributions. Technical Report: Lab. Stat. Prob., Toulouse, France (2004).

[50] P.L. Leung. An identity for the noncentral wishart distribution with application. Journal of Multivariate Analysis. vol. 48, no. 1. pp. 107–114 (1994)

[51] S.C. Liaw, H.G. Yeh, F.K. Hwang, and G.J. Chang. A simple and direct derivation for the number of noncrossing partitions. Proceedings of the American Mathematical Society. vol. 126, no. 6. pp. 1579–1581 (1998).

[52] E.H. Lieb. Convex Trace Functions and the Wigner–Yanase–Dyson Conjecture. Advances in Math. vol. 11. pp. 267–288 (1973).

[53] G. Lin and J. Stoyanov. On the Moment Determinacy of Products of Non-identically Distributed Random Variables. arXiv e-print, arXiv:1406.1654 (2014).

[54] F. Lust-Piquard. Inégalités de Khintchine dans C_p $(1 < p < \infty)$. C.R. Acad. Sci. Paris. Sér. I Math. vol. 303, no. 7. pp. 289–292 (1986).

[55] F. Lust-Piquard and G. Pisier. Noncommutative Khintchine and Paley inequalities. Arkiv för Matematik. vol. 29, no. 1–2. pp. 241–260 (1991).

[56] V.A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik. vol. 114, no. 4. pp. 507–536 (1967).

[57] H. Markowitz. Portfolio Selection: Efficient Diversification of Investments. John Wiley and Sons, New York (1959).

[58] E. Mayerhofer, O. Pfaffel, and R. Stelzer. On strong solutions for positive definite jump-diffusions. Stochastic Processes and Their Applications. vol. 121, no. 9. pp. 2072–2086 (2011).

[59] M.L. Mehta. Random Matrices. 3rd Edition. Elsevier/Academic Press, Amsterdam (2004).

[60] J.A. Mingo and R. Speicher. Free Probability and Random Matrices. Springer (2016).

[61] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. The Quarterly Journal of Mathematics. vol. 11, no. 1. pp. 50–59 (1960).

[62] R.J. Muirhead. Aspects of Multivariate Statistical Theory. Wiley, New York (1982).

[63] K.P. Murphy. Machine Learning, A Probabilistic Perspective. Adaptive Computation and Machine Learning, The MIT Press (2012).

[64] K.P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report (2007).

[65] A. Nica and R. Speicher. Lectures on the combinatorics of free probability. Cambridge University Press (2006).

[66] A. Nica, D. Shlyakhtenko, and R. Speicher. Operator-valued distributions. I. Characterizations of freeness, International Mathematics Research Notices, pp. 1509–1538 (2002).

[67] H.J. Park and E. Ayanoglu. An upper bound to the marginal PDF of the ordered eigenvalues of Wishart matrices and its application to MIMO diversity analysis. In Proc. of the 2010 IEEE International Conference on Communications, Cape Town, South Africa (July, 2010).

[68] K. Phillips. R-Functions to Symbolically Compute the Central Moments of the Multivariate Normal Distribution. Journal of Statistical Software, Code Snippets. vol. 33, no. 1. pp. 1–14 (2010).

[69] F. Pourkamali-Anaraki. Estimation of the sample covariance matrix from compressive measurements. IET Signal Processing, vol. 10, no. 9, 1089–1095 (2016).

[70] J. Quaintance and H.W. Gould. Combinatorial Identities for Stirling Numbers. The unpublished notes of H.W. Gould. World Scientific Publishing, Singapore (2016).

[71] J. Quaintance and H.W. Gould. Table for Fundamentals of Series: Table III: Basic Algebraic Techniques. From the seven unpublished manuscripts of H.W. Gould, edited and compiled by J. Quaintance (2010).

[72] C.E. Rasmussen and C.K.I. Williams. Gaussian Processes for Machine Learning. The MIT Press (2006).

[73] C. Redelmeier and I. Emily. Genus expansion for real Wishart matrices. J. Theoret. Probab., vol. 24, pp. 1044–1062 (2011).

[74] J. Riordan. Enumeration of plane trees by branches and endpoints. Journal of Combinatorial Theory, Series A. vol. 19, no. 2. pp. 214–222 (1975).

[75] M. Rudelson. Random vectors in the isotropic position. Journal of Functional Analysis. vol. 164, no. 1. pp. 60–72 (1999).

[76] L.W. Shapiro. A Catalan triangle. Discrete Math. vol. 14, no. 1. pp. 83–90 (1976).

[77] H. Shin, M.Z. Win, J.H. Lee, and M. Chiani. On the capacity of doubly correlated MIMO channels. IEEE Transactions on Wireless Communications. vol. 5, no. 8. pp. 2253–2265 (2006).

[78] A.N. Shiryaev. Probability. Springer-Verlag, New York (1996).

[79] J. Stoyanov and G. Lin. Hardy's condition in the moment problem for probability distributions. Theory of Probability & Its Applications. vol. 57, no. 4. pp. 699–708 (2013).

[80] J.W. Silverstein and P.L. Combettes. Signal detection via spectral theory of large dimensional random matrices. IEEE Transactions on Signal Processing. vol. 40, no. 8. pp. 2100–2105 (1992).

[81] R. Simion. Noncrossing partitions. Discrete Mathematics. vol. 217, no. 1–3. pp. 367–409 (2000).

[82] R. Simion. Combinatorial statistics on non-crossing partitions. Journal of Combinatorial Theory, Series A. vol. 66, no. 2. pp. 270–301 (1994).

[83] R. Simion and D. Ullman. On the structure of the lattice of noncrossing partitions. Discrete Mathematics. vol. 98, no. 3. pp. 193–206 (1991).

[84] N.J.A. Sloane. The on-line encyclopaedia of integer sequences: https://oeis.org/.

[85] R.P. Stanley. Enumerative Combinatorics, vol. 1 and 2. Cambridge University Press (1999).

[86] R. Speicher. Combinatorial theory of the free product with amalgamation and operator-valued free probability theory, Mem. Amer. Math. Soc., vol. 132, no. 627, pp. x+88 (1998).

[87] R. Speicher and C. Vargas. Free deterministic equivalents, rectangular random matrix models, and operator- valued free probability theory Random Matrices: Theory and Applications (2012).

[88] T. Sugiyama. Percentile points of the largest latent root of a matrix and power calculation for testing $\Sigma = I$. Journal of the Japan Statistical Society, Japanese Issue. vol. 3, no. 1. pp. 1–8 (1973).

[89] T. Tao. Topics in random matrix theory. American Mathematical Society, Providence, Rhode Island (2012).

[90] S. Thorbjornsen. Mixed moments of Voiculescu's Gaussian random matrices, Journal of Functional Analysis, vol. 176, pp. 213–246 (2000).

[91] J.Y. Tourneret, A. Ferrari, and G. Letac. The noncentral Wishart distribution: Properties and application to speckle imaging. In Proceedings of the 2005 IEEE/SP 13th Workshop on Statistical Signal Processing, Bordeaux, France (July, 2005).

[92] D.S. Tracy and S.A Sultan. Higher Order Moments of Multivariate Normal Distribution Using Matrix Derivatives. Stochastic Analysis and Applications. vol. 11, no. 3. pp. 337–348 (1993).

[93] J.A. Tropp. User-Friendly Tail Bounds for Sums of Random Matrices. Foundations of Computational Mathematics. vol. 12, no. 4. pp. 389–434 (2012).

[94] J.A. Tropp. An Introduction to Matrix Concentration Inequalities. Foundations and Trends in Machine Learning. vol. 8, no. 1-2. pp. 1–230 (2015).

[95] D.V. Voiculescu. Operations on certain non-commutative operator-valued random variables, Asterisque, no. 232, pp. 243–275 (1995).

[96] H. Weyl. The laws of asymptotic distribution of the eigenvalues of linear partial differential equation. Math. Ann. vol. 71. pp. 441–479 (1912).

[97] G.C. Wick. The evaluation of the collision matrix. Physical Review. vol. 80, no. 2, pp. 268–272 (1950).

[98] E. Wigner. Characteristic Vectors of Bordered Matrices With Infinite Dimensions. The Annals of Mathematics. vol. 62, no. 3. pp. 548–564 (1955).

[99] J.H. Winters. On the capacity of radio communication systems with diversity in Rayleigh fading environment. IEEE Journal on Selected Areas in Communications. vol. 5, no. 5. pp. 871–878 (1987).

[100] J. Wishart. The generalized product moment distribution in samples from a normal multivariate population. Biometrika, vol. 20A, no. 1/2. pp. 32–52 (1928).

[101] F. Yano and H. Yoshida. Some set partition statistics in non-crossing partitions and generating functions. Discrete Mathematics. vol. 307, no. 24. pp. 3147–3160 (2007).

[102] Y.Q. Yin. Limiting spectral distribution for a class of random matrices. Journal of Multivariate Analysis. vol. 20, no. 1. pp. 50–68 (1986).

[103] Y.Q. Yin and P.R. Krishnaiah. A limit theorem for the eigenvalues of product of two random variables. Journal of Multivariate Analysis. vol. 13, no. 4. pp. 489–507 (1983).

[104] A. Zanella, M. Chiani, and M.Z. Win. On the marginal distribution of the eigenvalues of Wishart matrices. IEEE Transactions on Communications. vol. 57, no. 4. pp. 1050–1060 (2009).