Research Statement

Jonathan Wells

October 2021

My research lies in the intersection of Probability, Statisitics, and Mathematical Physics, in the field of Random Matrix Theory. Currently I study the solvability of β-ensembles of random matrices for integer values of β beyond 1, 2 and 4, along with their applications to statistical theory.[1]

Background

Random matrix theory is the study of the eigenvalue statistics of ensembles of matrices, which are a collection of (typically) square matrices along with a probability measure defined on this set, which in turn, induces a probability measure on the eigenvalues of the matrix. Although a rich field of study on its own, random matrix theory also enjoys rather abundant application throughout mathematics, statistics and physics, since the eigenvalue statistics of many matrix ensembles can be used to model a wide variety of phenomena, including the statistics of discrete energy levels in atomic spectra [9], the dynamics of a multi-particle system interacting via a repulsive force [2], and the estimates of the covariance matrix for a population vector [10]. The widespread applicability of random matrices evinces a universal paradigm—a collection of theorems akin to the classical Central Limit Theorem [5]. But the utility of such theorems depends on an available supply of *solvable* ensembles in each universality class—collections of matrices for which the densities of eigenvalues can be expressed in terms of 'known' functions whose properties and asymptotics are well-studied.

The β-ensembles are one such collection, and are composed of random matrices whose eigenvalue densities take a common form, indexed by a non-negative, real parameter β:

$$\rho_N(x_1, \ldots, x_N) = \frac{1}{Z_N(\beta) N!} \prod_{i<j} |x_j - x_i|^\beta \prod_i w(x_i) \tag{1}$$

where $w(x)$ is a specified probability density (often Gaussian, but not necessarily) and where $Z_N(\beta)$ denotes the *partition function* of β, and is the normalizing constant required for ρ_N to be a probability density function. It turns out that the classic β-ensembles ($\beta = 1, 2, 4$ with $w(x) = \exp(-\frac{x^2}{2})$ describe the eigenvalue distributions of Hermitian matrices with real, complex, or quaternionic Gaussian entries [5].

The $\beta = 2$ ensemble is an example of a *determinantal point process* (meaning the densities can be obtained as the determinant of a matrix whose entries are given by a two variable kernel function), while the $\beta = 1, 4$ ensembles are examples of *Pfaffian point processes* (meaning the densities can be obtained as the Pfaffian of an antisymmetric matrix whose entries are given by a kernel function). For present purposes, it suffices to define the Pfaffian as a square root of the determinant (although it should be noted that there are other, equivalent definitions which turn out to be more amenable to algebraic manipulation).

Of fundamental concern in the theory of random matrices is the behavior of eigenvalue statistics of matrix ensembles as $N \to \infty$. The immediate advantage of aforementioned determinantal and Pfaffian formulations for the density functions is that the fundamental characteristics of the eigenvalues are encoded in the kernel functions, which do not essentially increase in complexity as N grows large, and which can be expressed as a sum whose asymptomatics are well-understood.

Consider, for example, the sample covariance matrix formed from a large sample of random vectors, and suppose we are interested in the eigendecomposition of this matrix (for the purposes of principal component analysis, for example). Assuming that the eigenvalue for this random matrix are approximately distributed according to the distribution of eigenvalues for a β-ensemble of matrices (which, by the universality paradigm

[1] What follows is a concise statement of my research interests written for a general mathematical and statistical audience. I am happy to provide a detailed research statement upon request.

described above, holds with remarkable frequency), then the asymptotic distribution for the β-ensemble eigenvalues also well-approximates the eigenvalue distribution for the covariance matrix. But the former is eminently accessible, thanks to the determinantal/Pfaffian formulation.

More generally, β-ensembles for values of β not equal to 1, 2, or 4 may also possess analogous determinant/Pfaffian kernel formulations. My current research uses tools from the exterior and shuffle algebras in order to unify the structure of the $\beta = 1, 2, 4$ ensembles, with eventual goal of representing β-ensembles as *Hyperpfaffian point processes* when β is an arbitrary square integer.

Current Work

In the classical cases ($\beta = 1, 2, 4$), the first step to rewriting the density functions as a determinant/Pfaffian is to observe that the partition function $Z_N(\beta)$ is itself a determinant/Pfaffian. Then, using the Cauchy-Binet Formula and a Sylvester Identity for matrices, it can be shown that the generating function for the density functions has determinantal/Pfaffian coefficients.

In order to extend this result more generally to the case when β is a square integer, it is helpful to recast the problem in a more algebraic light. Of chief concern are the exterior and shuffle algebras. The former is obtained from the tensor algebra by imposing the relation $a \otimes b = -b \otimes a$, while the latter is obtained by endowing the tensor algebra with an additional multiplication, where the product of two tensors is taken to be the sum of all tensors obtained by interlacing the two.

- In [6], I show that when β is a square integer, techniques in the shuffle algebra can be used to write the partition function as the Hyperpfaffian of an antisymmetric tensor, which is a higher-dimensional analogue of the Pfaffian of an antisymmetric matrix. This partition function can then be used to model the electrostatics distribution of charged particles that are confined to a line, interact with pairwise logarithmic repulsion, and are in thermal equilibrium with temperature β^{-1}.

- More generally, in [8], we show that an analogous result holds when β is allowed to depend on the choice of indices (i, j) in the product in Equation 1. As above, this partition function models the electrostatic distribution of charged particles with logarithmic repulsion, but where particles are allowed to have distinct charges.

The existence of this Hyperpfaffian representation for the partition function suggests that the eigenvalue density functions also may admit a Hyperpfaffian representation (just as the observation that the partition function is a Pfaffian when $\beta = 1, 4$ is the first step in showing that the density functions can be written as Pfaffians). And indeed, there are Hyperpfaffian analogues for the algebraic tools used to show that the $\beta = 1, 4$ density functions are Pfaffians. Unfortunately, it appears that the Sylvester Identity trick (a fundamental maneuver for the Pfaffian case) applies only to antisymmetric tensors which can be written in a certain 'invertible' form after a suitable change of basis, and this is not possible for a general antisymmetric tensor.

Directions for Future Research

The 'invertible' element obstacle described above is not insurmountable problem, however, since the antisymmetric tensors that arise from the β-ensemble model are highly non-generic. Indeed, I have recently focused on several directions for further research:

1. Since the the partition function $Z_N(\beta)$ depends only on the Hyperpfaffian of an L-tensor (and not the L-tensor itself), the particular L-tensors arising from the β-ensemble may be replaced in a deterministic fashion by other antisymmetric tensors which can be 'inverted'.

2. Although the general β-ensemble may not readily admit a Hyperpfaffian representation, it is possible that the circular β-ensembles (which historically are more tractable than the classic Gaussian ensembles) give rise to antisymmetric tensors which can be directly computed and chosen in such a way to allow application of the Sylvester identity.

3. Alternatively, there may be other point processes on $\mathbb{R}$ (either naturally arising or contrived for this purpose) whose associated antisymmetric tensors are reasonably well-behaved. It may then be possible to approximate the β-ensemble point processes using these 'Hyperpfaffian' models.

A first step towards investigating the first and second topic above involves determining necessary and sufficient conditions for the 'invertibility' of an antisymmetric tensor. This will likely need to be done both for the case of generic tensors, as as for the cases when the tensor coefficients arise from the β-ensemble models. While the formal computations involving particular sums of powers of these antisymmetric tensors are intractably cumbersome, much of the complexity can be shed by instead simulating 'values' of these sums for large samples of randomly selected antisymmetric tensors and analyzing the results using appropriate data visualization techniques. Essentially, we may be able to understand the distribution of these tensor powers via classic Monte Carlo methods. During the summer of 2020, I worked alongside J. Li, a Reed student supported by the Reed College Science Research Fellowship, on a project investigating the dynamics and stochastic behavior of these tensor powers. Our results are summarized in a forthcoming paper, currently in preparation [7].

Based on prior work, in order to address the third topic above, it seems fruitful to further investigate statistical inference procedures for distinguishing between non-homogeneous Poisson point processes and determinantal and Pfaffian point processes, as well as procedures for estimating the parameters of β-ensemble distributed points. Ultimately, my goal is to obtain, visualize, and analyze data from β-ensemble point processes when β is square integer in order to determine the essential characteristics for 'Hyperpfaffian point processes.'

Undergraduate Reading and Research Projects

During my previous two years as a Visiting Assistant Professor of Statistics at Reed College, I advised seven year-long senior thesis projects in probability, statistics, computer science and economics, and am currently supervising an additional four projects on topics arising from the intersection of my research experience and the students' own academic interests. Previous to this, as a graduate student at the University of Oregon, I supervised six multi-term undergraduate reading and research projects, in conjunction with the UO Association for Women in Mathematics Undergraduate Reading Program and the University of Oregon Directed Reading program. Brief descriptions of a selection of these projects are given below:

1. One project analyzes optimal decision-making and long-term trends in a series of competitive games with asymmetric roles in which players are allowed to modify strategy and choice of role between games. The student modeled these series a non-linear discrete time stochastic processes and investigated the mixing time, stationary and ergodic properties, and robustness of these processes.

2. Another project was less conventional, and arose out of the student's interest in Sudoku puzzles. The set of 'solved' Sudoku puzzles (9×9 grids filled with the numbers 1 through 9 so each row, column, and principle 3×3 sub-grid contains each value exactly once) can be viewed as a particular subset of the collection of doubly stochastic matrices. Using techniques from statistics, group theory, and random matrices, the student and I developed an algorithm for generating a sufficiently rich ensemble of Sudoku matrices. Using this algorithm, we conducted an empirical investigation of the eigenvalue and trace behavior for a randomly chosen Sudoku matrix.

3. A second project investigated the collaboration and citation networks among R&D firms and universities, explored the graph-theoretic characteristics of these networks in order to predict the impact of research innovations and advancement in specific disciplines. The student used Markov chain techniques to perform exploratory data analysis on these networks and used the results to create a multivariate regression model relating citation connectivity to degree of innovation.

4. A third project studied the Tracy-Widom distribution, which arises in random matrix theory as the limiting law for the largest eigenvalue of a Hermitian matrix with independent Gaussian entries. In particular, after an in-depth literature review, the student codified procedures for performing hypothesis testing with the Tracy-Widom distribution as part of a novel non-parametric test for independence.

5. A recent project investigated latent structure random graphs, where rather than assuming that vertices are connected via an edge independently with constant probability (in the manner of the Erdös-Réyni random graph), instead vertices are connected with heterogeneous probability that depends on the location of those points in a corresponding latent space. Much of the initial structure of the project was based on the recent paper by A. Athreya et al. [1].

In addition to the current and past projects outlined above, I have given careful consideration to several topics in mathematics, statistics, and mathematical physics related to my research program which would be amenable to undergraduate collaboration or independent study. A small selection of these topics follow.

1. Statistical inference procedures for Poisson point processes have been relatively well-studied, but analogous procedures for determinantal and Pfaffian processes are considerably less developed. I would like to work with a student with significant theoretical statistics background to develop criteria for identifying real-world phenemona which can be modeled as determinantal/Pfaffian point processes, as well creating hypothesis testing procedures for distinguishing between these types of processes.

2. Under the Boltzman statistics paradigm, a wealth of information about a particle system's behavior is encoded in the partition function (or the ensemble average) But in certain cases when the system behaves like a log-Coulomb gas, this partition function can be represented in a particularly nice algebraic form given in terms of the particle positions. I would like to work with a student with algebra or combinatorics background to explore some of the combinatorial identities that arise in these computations.

3. A classic result due to Karlin and McGregor [4] shows that for certain non-intersecting random walks, the distribution of points midway along a path can be represented by a determinantal process. A recent result by Garrod et al.[3] demonstrates that a class of annihilating/coalescing random walks can represented as a Pfaffian point process. I would like to work with a student with significant probability and stochastic process background to determine whether random walk models exist for arbitrary β-ensemble-like covariance structure.

References

[1] A. Athreya, M. Tang, Y. Park, and C. Priebe. On estimation and inference in latent structure random graphs. *Statistical Science*, 36, 02 2021.

[2] P. J. Forrester. *Log-Gases and Random Matrices*. London Mathematical Society Monographs. Princeton University Press, 2010.

[3] B. Garrod, M. Poplavskyi, R. P. Tribe, and O. V. Zaboronski. Examples of interacting particle systems on z as pfaffian point processes: Annihilating and coalescing random walks. *Annales Henri Poincare*, 19(12):3635–3662, 2018.

[4] S. Karlin and J. McGregor. The differential equations of birth-and-death processes, and the stieltjes moment problem. 1957.

[5] M. L. Mehta. *Random Matrices*. Elsevier/Academic Press, 2004.

[6] J. Wells. *On the Solvability of Beta-Ensembles when Beta is a Square Integer*. PhD thesis, University of Oregon, 2019.

[7] J. Wells and J. Li. A collection of hyperpfaffian identities in the exterior algebra. 2020. In preparation.

[8] J. Wells and E. Wolff. The partition function of log-gases with multiple odd charges. Submitted 2021. arXiv:2105.14378.

[9] E. P. Wigner. Characteristic vectors of bordered matrices with infinite dimension. *Annals of Mathematics*, 62, 1955.

[10] J. Wishart. The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, 20 A, 1928.