Free Probability, Sample Covariance Matrices and Stochastic Eigen-Inference

Alan Edelman
Department of Mathematics,
Computer Science and AI Laboratories.
E-mail: edelman@math.mit.edu
N. Raj Rao
Deparment of EECS,
MIT/WHOI Joint Program.
E-mail: raj@mit.edu

***Abstract*— Free probability provides tools and techniques for studying the spectra of large Hermitian random matrices. These stochastic eigen-analysis techniques have been invaluable in providing insight into the structure of sample covariance matrices. We briefly outline how these techniques can be used to analytically predict the spectrum of large sample covariance matrices. We discuss how these eigen-analysis tools can be used to develop eigen-inference methodologies.**

***Index Terms*— Free probability, random matrices, stochastic eigen-inference, rank estimation, principal components analysis.**

I. Introduction

The search for structure characterizes the nature of research in science and engineering. Mathematicians look for structure in difficult problems – discovering or even imposing a structure on the problem allows them to analyze a previously intractable problem. Engineers look to use this structure to gain insights into their algorithms and hopefully exploit the structure to improve the design. This article describes how the operator algebraic invention of free probability provides us with fresh insights into sample covariance matrices. We briefly mention an application of these techniques to an eigen-inference problem (rank estimation) that dramatically outperforms solutions found in the literature.

II. Free Probability and Random Matrices

A. Classical probability

We begin with a viewpoint on the familiar "classical" probability. Suppose we are given a random variable a whose probability distribution is a compactly supported probability measure on $\mathbb{R}$, which we denote by μ_a. The moments of the random variable a, denoted by $\varphi(a^n)$, are given by:

$$\varphi(a^n) = \int_{\mathbb{R}} t^n d\mu_a(t). \tag{1}$$

More generally, if we are given the probability densities μ_a and μ_b for *independent* random variables, a and b, respectively we can compute the moments of $a+b$ and ab from the moments of a and b. Specifically, our ability to do so is based on the fact that:

$$\varphi(a^{n_1}b^{m_1}\ldots a^{n_k}b^{m_k}) = \varphi(a^{n_1+\ldots n_k}b^{m_1+\ldots+m_k}) \tag{2}$$

since a and b commute and are independent. In particular, the distribution for $a+b$, when a and b are independent, is simply the convolution of the measures μ_a and μ_b. A more familiar way of restating this result is that the Fourier transform of the probability measure of the sum of two independent random variables is the product of the Fourier transforms of the individual probability measures.

B. Free probability

We adopt a similar viewpoint on free probability using large random matrices as an example of "free" random variables. Throughout this paper, let $\mathbf{A}_N$ be an $N \times N$ symmetric (or Hermitian) random matrix with real eigenvalues. The probability measure on the set of its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ (counted with multiplicities) is given by:

$$\mu_{A_N} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i}. \tag{3}$$

We are interested in the limiting spectral measure μ_A as $N \to \infty$ which, when compactly supported, is uniquely characterized by the moments computed as in (1). We refer to $\mathbf{A}$ as an element of the "algebra" with probability measure μ_A and moments $\varphi(A^n)$.

Suppose we are now given two random matrices $\mathbf{A}_N$ and $\mathbf{B}_N$ with limiting probability measures μ_A and μ_B, we would like to compute the limiting probability measures for $\mathbf{A}_N + \mathbf{B}_N$ and $\mathbf{A}_N\mathbf{B}_N$ in terms of the

moments of μ_A and μ_B. It turns out that the appropriate structure, analogous to independence for "classical" random variables, that we need to impose on $\mathbf{A}_N$ and $\mathbf{B}_N$ to be able to compute these measures is "freeness".

It is worth noting, that since $\mathbf{A}$ and $\mathbf{B}$ do not commute we are operating in the realm of non-commutative algebra. Since all possible products of $\mathbf{A}$ and $\mathbf{B}$ are allowed we have the "free" product, i.e., all words in $\mathbf{A}$ and $\mathbf{B}$ are allowed. (We recall that this is precisely the definition of the free product in algebra.) The theory of free probability allows us to compute the moments of these products. The connection with random matrices comes in because a pair of random matrices $\mathbf{A}_N$ and $\mathbf{B}_N$ are asymptotically free, i.e., in the limit of $N \to \infty$ so long as at least one of $\mathbf{A}_N$ or $\mathbf{B}_N$ has what amounts to eigenvectors that are uniformly distributed with Haar measure. This result is stated more precisely in [6].

As was the case with independence for "classical" random variables, "freeness" is the structure needed to be able to compute mixed moments of the form $\varphi(\mathbf{A}^{n_1}\mathbf{B}^{m_1}\ldots\mathbf{A}^{n_k}\mathbf{B}^{m_k})$. We note that the restriction that A and B do not commute so that in general,

$$\varphi(\mathbf{A}^{n_1}\mathbf{B}^{m_1}\ldots\mathbf{A}^{n_k}\mathbf{B}^{m_k}) \neq \varphi(\mathbf{A}^{n_1+\ldots n_k}\mathbf{B}^{m_1+\ldots+m_k}). \tag{4}$$

is embedded into the definition of "freeness" when it was invented by Voiculescu [6] in the context of his studies on operator algebras. Though the condition for establishing "freeness" between a pair of random matrices, as described in [6], is quite technical and appears abstract, it naturally arises many practical scenarios as detailed in Section III.

C. Free Multiplicative Convolution

When $\mathbf{A}_N$ and $\mathbf{B}_N$ are asymptotically free, the (limiting) probability measure for random matrices of the form $\mathbf{A}_N\mathbf{B}_N$ (by which we really mean the self-adjoint matrix formed as $\mathbf{A}_N^{1/2}\mathbf{B}_N\mathbf{A}_N^{1/2}$) is given by the *free multiplicative convolution* [6] of the probability measures μ_A and μ_B and is written as $\mu_{AB} = \mu_A \boxtimes \mu_B$. The algorithm for computing μ_{AB} is given below.

Step 1: Compute the Cauchy transforms, $G_A(z)$ and $G_B(z)$ for the probability measures μ_A and μ_B respectively. For a probability measure μ on $\mathbb{R}$, the Cauchy transform is defined as:

$$G(z) = \int_{\mathbb{R}} \frac{1}{z-t} d\mu(t). \tag{5}$$

This is an analytic function in the upper complex half-plane. We can recover the probability measure from the Cauchy transform by the Stieltjes inversion theorem which says that:

$$d\mu(t) = -\frac{1}{\pi}\lim_{\epsilon\to 0} \Im G(t+i\epsilon), \tag{6}$$

where the $\Im$ denotes the imaginary part of a complex number.

Step 2: Compute the ψ-transforms, $\psi_A(z)$ and $\psi_B(z)$. Given the Cauchy transform $G(z)$, the ψ-transform is given by:

$$\psi(z) = \frac{G(1/z)}{z} - 1 \tag{7}$$

Step 3: Compute the S-transforms, $S_A(z)$ and $S_B(z)$. The relationship between the ψ-transform and the S-transform of a random variable is given by:

$$S(z) = \frac{1+z}{z}\,\psi^{\langle -1\rangle}(z) \tag{8}$$

where $\psi^{\langle -1\rangle}(z)$ denotes the inverse under composition.

Step 4: The S-transform for the random variable $\mathbf{AB}$ is given by:

$$S_{AB}(z) = S_A(z)S_B(z). \tag{9}$$

Step 5: Compute $\psi_{AB}(z)$ from the relationship in (8).

Step 6: Compute the Cauchy transform, $G_{AB}(z)$ from the relationship in (7).

Step 7: Compute the probability measure μ_{AB} using the Stieltjes inversion theorem in (6).

D. Free Additive Convolution

When $\mathbf{A}_N$ and $\mathbf{B}_N$ are asymptotically free, the (limiting) probability measure for random matrices of the form $\mathbf{A}_N + \mathbf{B}_N$ is given by the *free additive convolution* of the probability measures μ_A and μ_B and is written as $\mu_{A+B} = \mu_A \boxplus \mu_B$. A similar algorithm in terms of the so-called R-transform exists for computing μ_{A+B} from μ_A and μ_B. See [6] for more details.

What we want to emphasize about the algorithms described in Sections (II-C) and (II-D) is simply that the convolution operators on the non-commutative algebra of large random matrices exists and can be computed. Symbolic computational tools are now available to perform these non-trivial computations efficiently. See [2], [3] for more details. These tools enable us to analyze the structure of sample covariance matrices and design algorithms that take advantage of this structure.

III. Stochastic Eigen-Analysis of Sample Covariance Matrices

Let $\mathbf{y}$ be a $n \times 1$ observation vector modeled as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{10}$$

where $\mathbf{A}$ is a $n \times L$ matrix, $\mathbf{x}$ is a $L \times 1$ "signal" vector and $\mathbf{w}$ is a $n \times 1$ "noise vector". This model appears frequently in many signal processing applications [5]. If $\mathbf{x}$ and $\mathbf{w}$ are modeled as independent Gaussian vectors with independent elements having zero mean and unit

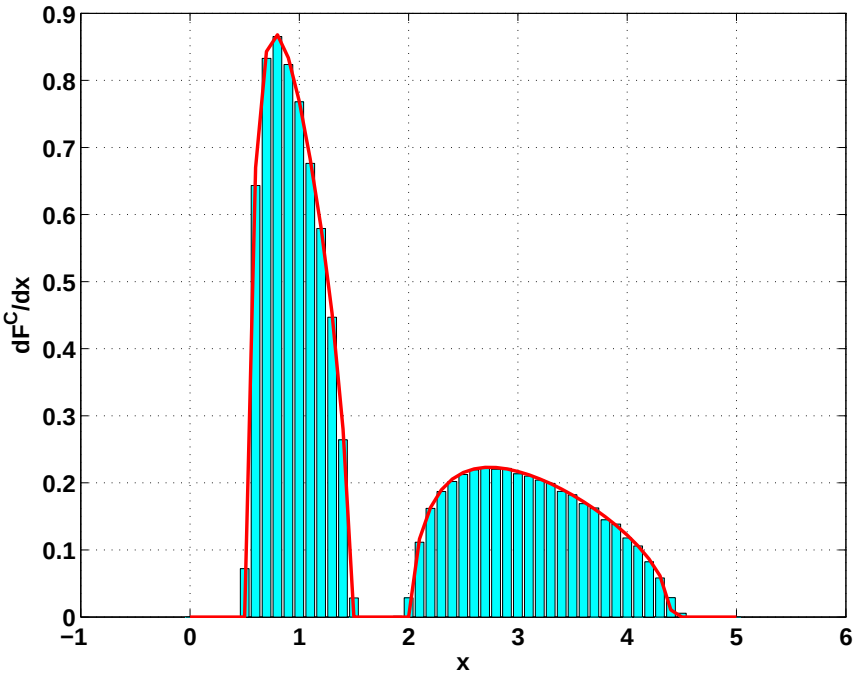

Fig. 1. The limiting spectral measure of a SCM (solid line) whose true measure is given in (16) for $P = 0.4$ and $\rho = 2$ compared with 1000 Monte-Carlo simulations for $n = 100$, $N = 1000$.

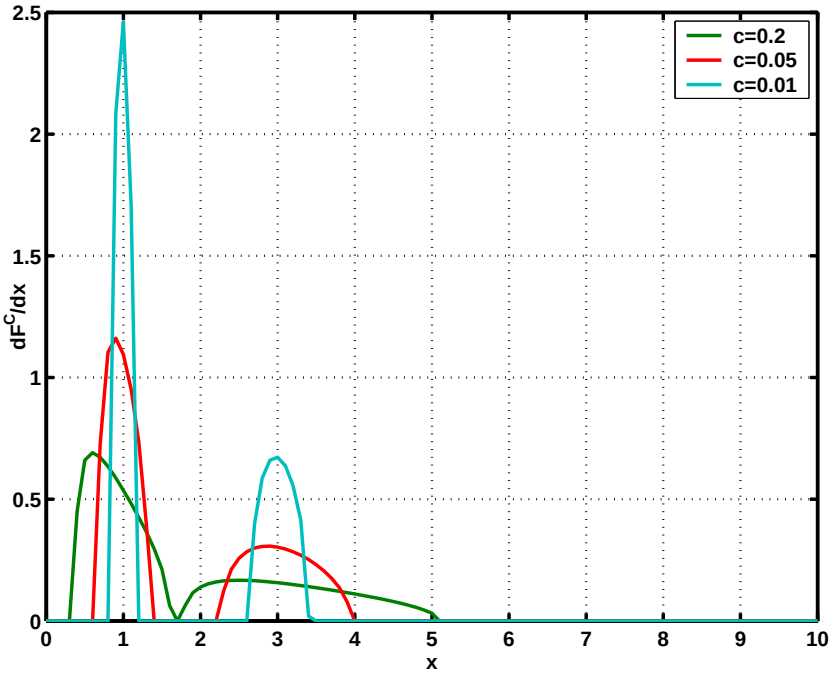

Fig. 2. The limiting spectral measure of a SCM whose true covariance matrix has measure (16) with $P = 0.4$ and $\rho = 2$, for different values of c. Note that as $c \to 0$, the blurring of the eigenvalues reduces.

variance (identity covariance), then $\mathbf{y}$ is a multivariate Gaussian with zero mean and covariance:

$$\mathbf{R} = E[\mathbf{y}\mathbf{y}^H] = \mathbf{A}\mathbf{A}^H + \mathbf{I}. \tag{11}$$

In most practical signal processing applications, the true covariance matrix is unknown. Instead, it is estimated from N independent observations ("snapshots") $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$ as:

$$\hat{\mathbf{R}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \mathbf{y}_i^H = \frac{1}{N}\mathbf{Y}_n \mathbf{Y}_n^H, \tag{12}$$

where $\mathbf{Y}_n = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N]$ is referred to as the "data matrix" and $\hat{\mathbf{R}}$ is the sample covariance matrix (SCM).

When n is fixed and $N \to \infty$, it is well-known the sample covariance matrix converges to the true covariance matrix. However, when both $n, N \to \infty$ with $n/N \to c > 0$, this is no longer true. Such a scenario is very relevant in practice where stationarity constraints limit the amount of data (N) that can be used to form the SCM. Free probability is an invaluable tool in such situations when attempting to understand the structure of the resulting sample covariance matrices.

We note first that the SCM can be rewritten as:

$$\hat{\mathbf{R}} = \mathbf{R}^{1/2}\mathbf{W}(c)\mathbf{R}^{1/2}. \tag{13}$$

Here $\mathbf{R}$ is the true covariance matrix. The matrix $\mathbf{W}(c) = (1/N)\mathbf{G}\mathbf{G}^H$ is the Wishart matrix formed from an $n \times N$ Gaussian random matrix with independent, identically distributed zero mean, unit variance elements. Once again, c is defined as the limit $n/N \to c > 0$ as $n, N \to \infty$.

Since the Wishart matrix thus formed has eigenvectors that are uniformly distributed with Haar measure, the matrices $\mathbf{R}$ and $\mathbf{W}(c)$ are asymptotically free! Hence the limiting probability measure $\mu_{\hat{R}}$ can be obtained using *free multiplicative convolution* as:

$$\mu_{\hat{R}} = \mu_R \boxtimes \mu_W \tag{14}$$

where μ_R is the limiting probability measure on the true covariance matrix $\mathbf{R}$ and μ_W is the Marčenko-Pastur density [1] given by:

$$\mu_W = \max\left(0, 1-\frac{1}{c}\right)\delta(x) + \frac{\sqrt{(x-b_-)(b_+-x)}}{2\pi x c} I_{[\,b_-, b_+]} \tag{15}$$

where $b_\pm = (1 \pm \sqrt{c})^2$ and $I_{[\,b_-,b_+]}$ equals 1 when $b_- \leq x \leq b_+$ and 0 otherwise.

IV. An Eigen-Inference Application

Let $\mathbf{A}\mathbf{A}^H$ in (10) have np of its eigenvalues of magnitude ρ and $n(1-p)$ of its eigenvalues of magnitude 0 where $p < 1$. This corresponds to $\mathbf{A}$ being an $n \times L$ matrix with $L < n$ with $p = L/n$ so that L of its singular values are of magnitude $\sqrt{\rho}$. Thus, as given by (11), the limiting spectral measure of $\mathbf{R}$ is simply:

$$\mu_R = p\,\delta(x-\rho-1) + (1-p)\,\delta(x-1). \tag{16}$$

Figure III compares the limiting spectral measure computed as in (14) with Monte-Carlo simulations. Figure III plots the limiting spectral measure as a function of c. Note that as $c \to 0$, we recover the measure in (16). The "blurring" in the eigenvalues of the SCM is because of insufficient sample support. When $c > 1$ then we are operating in a "snapshot deficient" scenario and the SCM is singular.

A. New rank estimation algorithm

Though the free probability results are exact when $n \to \infty$ the predictions are very accurate for $n \approx 10$ as well. If the example in (16) was a rank estimation algorithm where the objective is to estimate p and ρ then Figure III intuitively conveys why classical rank estimation algorithms such as [7] do not work as well as expected when there is insufficient data. Our perspective is that since free multiplicative convolution predicts the spectrum of the SCM that accurately we can use free multiplicative *deconvolution* to infer the parameters of the underlying covariance matrix model from a realization of the SCM! We are able to do this rather simply by "moment matching". The first three moments of the SCM can be analytically parameterized in terms of the unknown parameters p, ρ and the known parameter $c = n/N$ as:

$$\varphi(\hat{\mathbf{R}}) = 1 + p\rho \tag{17}$$

$$\varphi(\hat{\mathbf{R}}^2) = p\rho^2 + c + 1 + 2\,p\rho\,c + 2\,p\rho + cp^2\rho^2 \tag{18}$$

$$\begin{aligned}\varphi(\hat{\mathbf{R}}^3) = {} & 1 + 3\,c + c^2 + 3\,\rho^2 p + 3\,\rho^3 c p^2 + 3\,p\rho \\ & + 9\,p\rho\,c + 6\,p^2\rho^2 c + 3\,c\rho^2 p + 3\,p\rho\,c^2 \\ & + 3\,p^2\rho^2 c^2 + p^3\rho^3 c^2 + \rho^3 p\end{aligned} \tag{19}$$

Given an $n \times N$ observation matrix $\mathbf{Y}_n$, we can compute estimates of the first three moments as $\hat{\varphi}(\hat{\mathbf{R}}^k) = \frac{1}{n}\mathrm{tr}[(\frac{1}{N}\mathbf{Y}_n\mathbf{Y}_n^*)^k]$ for $k = 1, 2, 3$. Since we know $c = n/N$, we can estimate ρ, p by simply solving the nonlinear system of equations:

$$(\hat{\rho}, \hat{p}) \;=\; \arg\,\min_{(\rho,p)>0} \Big\| \sum_{k=1}^{3} \varphi(\hat{\mathbf{R}}^k) - \hat{\varphi}(\hat{\mathbf{R}}^k) \Big\|^2 \tag{20}$$

As $n, N \to \infty$ we expect the algorithm to perform well. It also performs well for finite n. Figure 3 compares the rank estimation performance of the new algorithm with the classical MDL/AIC based algorithm. The plots were generated over 2000 trials of an $n = 200$ system with $\rho = 1$, and $p = 0.5$ and different values of N. This implies that the true rank of the system is $n\,p = 100$. The bias of the rank estimation algorithm to be the

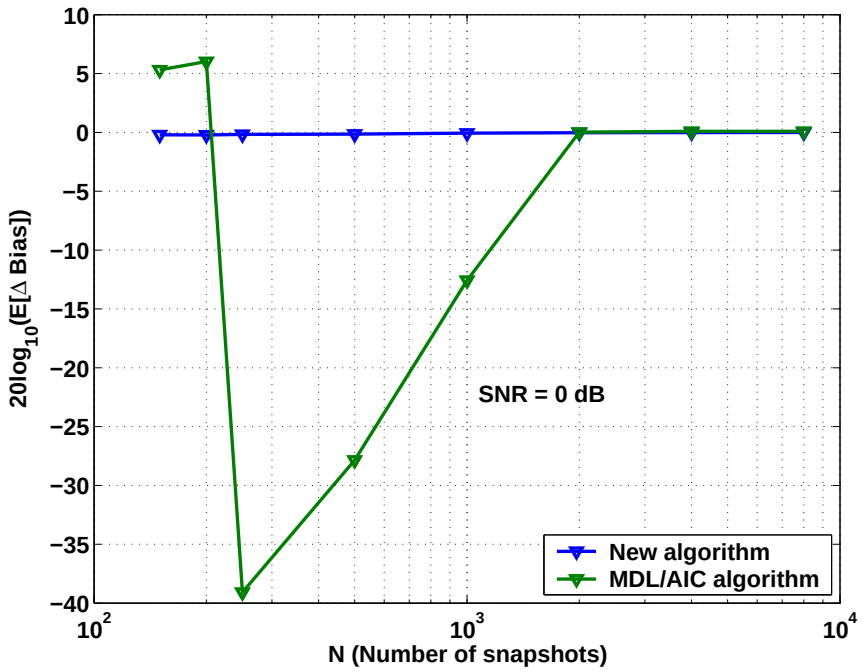

Fig. 3. (Relative) Bias in estimating rank of true covariance matrix: New algorithm vs. classical algorithm ($n = 200$).

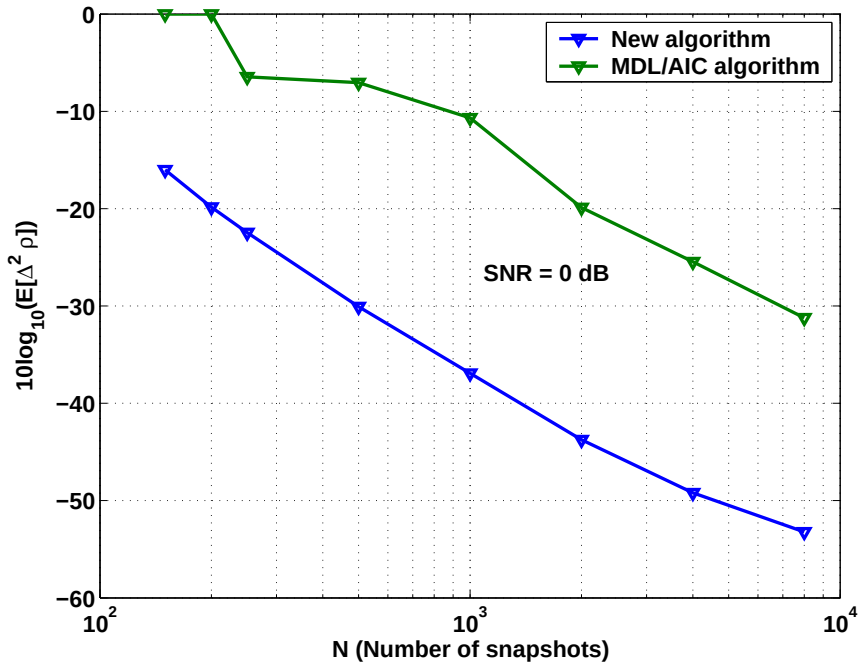

Fig. 4. Mean squared error in estimating ρ: New algorithm vs. classical algorithm ($n = 200$).

ratio of the estimated rank to the true rank. Hence 0 dB corresponds to zero rank estimation error and so on. As Figure 3 indicates, the new algorithm dramatically outperforms the classical algorithm and remains, on the average, within 1 dimension (i.e. < 0.2 dB) of the true rank even when in the snapshot deficient scenario, i.e., $N < n$! Additionally, the new rank estimation algorithm can be used to estimate ρ and p. Figure 4 compares the mean-squared estimation error for ρ for the new and the MDL/AIC algorithm respectively. Though the MDL/AIC estimation error is fairly small, the new algorithm, once again, performs significantly better, especially when $n \approx N$. See [4] for a generalization of this algorithm including a rigorous theoretical analysis of its estimation properties.

V. Summary

Free probability, which has deep connections to the studies of operator algebras, is an invaluable tool for describing the spectral of large sample covariance matrices. See [5] for applications to performance analysis. As the algebraic structure captured by free probability gets increasingly familiar, additional applications that exploit this structure to design improved algorithms (as in Section IV-A) are bound to emerge. This is yet another instance of how the search for structure in signal processing leads to new analytic insights, applications and directions for future research.

Acknowledgements

N. Raj Rao's work was supported by NSF Grant DMS-0411962. Alan Edelman's work was supported by SMA.

References

[1] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)*, 72 (114):507–536, 1967.

[2] N. R. Rao and A. Edelman. The polynomial method for random matrices. Preprint, 2005.

[3] N. Raj Rao. *Infinite random matrix theory for multi-channel signal processing*. PhD thesis, Massachusetts Institute of Technology, 2005.

[4] N. Raj Rao and A. Edelman. Eigen-inference from large Wishart matrices. preprint, 2005.

[5] A. M. Tulino and S. Verdú. Random matrices and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1), June 2004.

[6] D. V. Voiculescu, K. J. Dykema, and A. Nica. *Free random variables*. Amer. Math. Soc., Providence, RI, 1992.

[7] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):387–392, April 1985.