Homework 1

Marc Potters

Introduction to Random Matrix Theory
and its applications to data analysis

January 15, 2019
due January 25, 2019

Exercise 1. Warm-up

- Let $\mathbf{M}$ be a random real symmetric orthogonal matrix, that is an $N \times N$ matrix satisfying $\mathbf{M} = \mathbf{M}^\intercal = \mathbf{M}^{-1}$. Show that all the eigenvalues of $\mathbf{M}$ are ± 1.
- Let $\mathbf{X}$ be a Wigner matrix, i.e. an $N \times N$ real symmetric matrix whose diagonal and upper triangular entries are iid Gaussian random numbers with zero mean and variance σ^2/N. You can use $\mathbf{X} = \sigma/\sqrt{2N}(\mathbf{H} + \mathbf{H}^T)$ where $\mathbf{H}$ is a non-symmetric $N \times N$ matrix will with iid standard Gaussians.
- The matrix $\mathbf{E}$ will be $\mathbf{E} = \mathbf{M} + \mathbf{X}$. $\mathbf{E}$ can be thought of as a noisy version of $\mathbf{M}$. The goal of these exercise is to understand numerically how the matrix $\mathbf{E}$ is corrupted by the Wigner noise.
- The matrix $\mathbf{P}_+$ is defined as $\mathbf{P}_+ = \frac{1}{2}(\mathbf{M} + \mathbf{1}_N)$. Convince yourself that $\mathbf{P}_+$ is the projector onto the eigenspace of $\mathbf{M}$ with eigenvalue $+1$. Explain the effect of the matrix $\mathbf{P}_+$ on eigenvectors of $\mathbf{M}$.
- An easy way to generate a random matrix $\mathbf{M}$ is to generate a Wigner matrix (independent of $\mathbf{X}$), diagonalize it, replace every eigenvalue by its sign and reconstruct the matrix. The procedure does not depend on the σ used for the Wigner.
- Using the computer language of your choice, for a large value of N (as large as possible while keeping computing times below one minute), for a three interesting values of σ of your choice, do the following numerical analysis.

(a) Plot a histogram of the eigenvalues of $\mathbf{E}$, for a single sample first, and then for many samples (say 100).

(b) From your numerical analysis, in the large N limit, for what values of σ do you expect a non-zero density of eigenvalue near zero.

(c) For every normalized eigenvector $\mathbf{v}_i$ of $\mathbf{E}$, compute the norm of the vector $\mathbf{P}_+\mathbf{v}_i$. For a single sample, do a scatter plot of $|\mathbf{P}_+\mathbf{v}_i|^2$ vs λ_i (its eigenvalue). Turn your scatter plot into an approximate conditional expectation value (using an histogram) including data from may samples.

(d) Build an estimator $\Xi(\mathbf{E})$ of $\mathbf{M}$ using only data from $\mathbf{E}$. We want to minimise the error $e = \frac{1}{N}||(\Xi(\mathbf{E}) - \mathbf{M})||_{\mathrm{F}}^2$ where $||A||_{\mathrm{F}}^2 = \mathrm{Tr} AA^{\intercal}$. Consider first $\Xi_1(\mathbf{E}) = \mathbf{E}$ and then $\Xi_0(\mathbf{E}) = 0$. What is the error e of these two estimators. Try to build an ad-hod estimator $\Xi(\mathbf{E})$ that has a lower error e than these two.

(e) Show numerically that the eigenvalues of $\mathbf{M}$ are not iid. For each sample $\mathbf{M}$ rank its eigenvalues $\lambda_1 < \lambda_2 < \ldots < \lambda_N$. Consider the eigenvalue spacing $s_k = \lambda_k - \lambda_{k-1}$ for eigenvalues in the bulk $(.2N < k < .3N$ and $.7N < k < .8N)$. Make an histogram of $\{s_k\}$ including data form 100 samples. Make a 100 pseudo-iid samples: mix eigenvalues for 100 different samples and randomly choose N from the $100N$ possibilities, do not choose the same eigenvalue twice for a given pseudo-iid sample. For each pseudo-iid sample, compute s_k in the bulk and make an histogram of the values using data from all 100 pseudo-iid samples. (Bonus) Try to fit a exponential distribution to these two histograms. The iid should be well fitted by the exponential but not the original data (not iid).