Mathematical Methods in Data Science and Signal Processing

Homework Assignment 2

December 21, 2022

General instructions: Upload your solution text and code to Moodle via the dedicated submission box.

1. **Power iteration.** Generate a random symmetric matrix of size $n \times n$ with $n = 1000$. Implement the power method and compute the leading eigenvector of the matrix. Compute the ground truth leading eigenvector v of the matrix using existing implementations of numerical algorithms, i.e., `scipy.eigs.sparse.linalg.eigsh` in Python and `eigs` in MATLAB. Compute the reconstruction error as a function of the iteration, which is defined as

   $$\text{relative error}(i) = \min_{z \in \{\pm 1\}} \frac{\|z\hat{v}(i) - v\|}{\|v\|}, \tag{1}$$

   where $\hat{v}(i)$ is the i-th iteration estimate.

   Explain why we define the error as we do in (1). Plot the error as a function of i. Scale your results so as to obtain roughly a straight line. Justify your choice of scale. What's the slope of the error curve? Does it fit the theory?

2. **Diffusion maps.** In this question we work on a "window" signal, a signal x of length 50, whose first 10 entries are ones and the rest are zeros. We shall have $n = 2000$ observations $y_1, \ldots, y_n$ drawn from the model

   $$y_i = R_{\ell_i} x + \epsilon_i, \quad i = 1, \ldots, n. \tag{2}$$

   Here, ϵ_i is a Gaussian noise with zero mean and variance σ^2 and $R_{\ell_i} x$ is a random circular shift, namely, $(R_{\ell_i} x)[j] = x[j - \ell_i]$, with indices starting from zero and taken modulo n and $\{\ell_i\}_{i=1}^n$ are independently drawn from a uniform distribution on $\{0, \ldots, n-1\}$.

   a) Set $\sigma = 0$ (no noise). Implement the diffusion map algorithm with a Gaussian kernel having a standard deviation τ_g.

      i. Plot the embedding of the n observations onto a two-dimensional space. Explain the results.

      ii. Which standard deviation τ_g did you use? What happens if you repeat the same experiment with $\tau_g/10$? Explain the result.

   b) Choose at least three different illustrative non-zero noise levels σ^2, generate noisy data according to (2), and plot its diffusion map embedding into two-dimensional space. How does the noise affect the two-dimensional embedding?

3. **Convex relaxation of max-cut.** Generate a graph with 40 vertices with 20 vertices belonging to set A and 20 vertices belonging to set B. For simplicity, you can assume that vertices $\{1, \ldots, 20\}$ are in A and vertices $\{21, \ldots, 40\}$ are in B. Now, generate a random graph according to the following rule. There is an edge between a pair of vertices with probability p if they belong to the same set, and with probability $1 - p$ if they are in different sets. A typical adjacency matrix for $p = 0.1$ appears in the figure below.

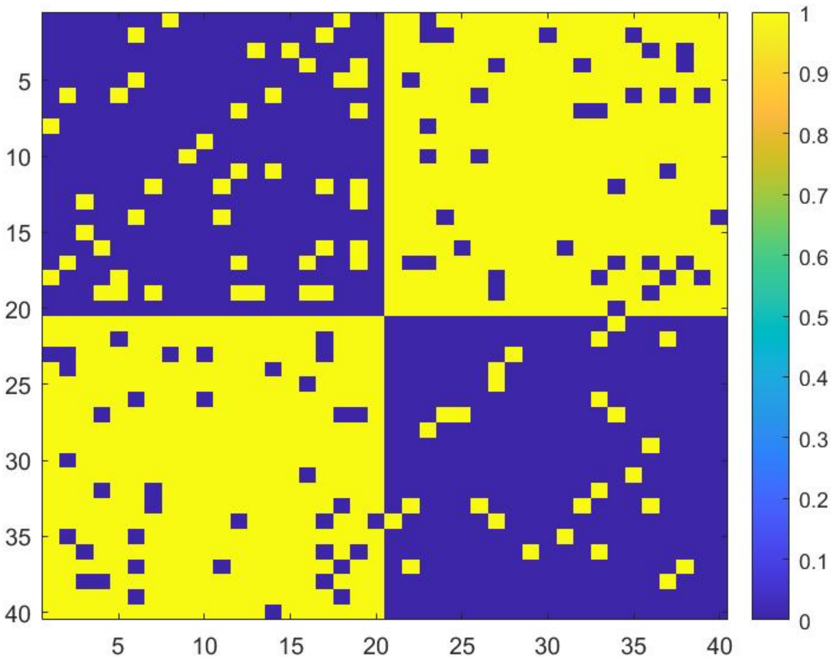

Figure 1: An example of an adjacency matrix with $p = 0.1$.

We wish to cluster the vertices into the two sets from the graph by finding the maximal cut. Implement the convex relaxation of max-cut using convex solvers, such as CVX for MATLAB or CVXPY for Python. Run this experiment 50 times for each value of p, for p's ranging from 0.1 to 0.5. Plot the average clustering error over the 50 trials as a function p. Here, the clustering error is the number of misclassified vertices divided by the total number of vertices (40). Succinctly describe in your own works the results you obtained.

4. **Synchronization.** Draw $n = 100$ angles $\theta := (\theta_1, \ldots, \theta_n)$ uniformly at random from the interval $[0, 2\pi)$. Let $\mathbf{h} = \left(e^{i\theta_1}, \ldots, e^{i\theta_n}\right)^\top \in \mathbb{C}^n$ and define the rank-one Hermitian matrix $\mathbf{H} = \mathbf{h}\mathbf{h}^* \in \mathbb{C}^{n\times n}$. Now, corrupt the matrix $\mathbf{H}$ according to the "outliers model", as follows. Let $j = 2, \ldots, n$ and $k = j+1, \ldots, n$ be indices of the upper triangular part of $\mathbf{H}$ strictly above the diagonal. For every such (j, k), in probability p replace the true (j, k)-th entry $\mathbf{H}_{j,k}$ with $e^{i\alpha}$, where α was sampled uniformly from $[0, 2\pi)$ (the "outlier"). Note that if you end up changing $\mathbf{H}_{j,k}$, you must also modify $\mathbf{H}_{k,j}$ appropriately so that $\mathbf{H}$ remains Hermitian.

   Now, estimate the rotations from corrupted matrices for p's ranging from 0 to 0.5. For each fixed p, conduct 50 trials. For each p and each trial, estimate the rotations using two methods, the spectral method and SDP. For each of the two methods, plot the average error over the 50 trials as a function of p.

   Recall that the estimate is defined up to a global angular shift, that is, hopefully $\mathbf{h} \approx \hat{\mathbf{h}} \cdot e^{i\theta_{\text{al}}}$ where $\hat{\mathbf{h}} \in \mathbb{C}^n$ are the estimated rotations and $\theta_{\text{al}} \in [0, 2\pi)$ is the alignment angle. Thus, the solution should be aligned with the ground-truth $\mathbf{h}$ before measuring the estimation error. Prove that

$$e^{i\theta_{\text{al}}} = \frac{\hat{\mathbf{h}}^*\mathbf{h}}{\left|\hat{\mathbf{h}}^*\mathbf{h}\right|}.$$

   and provide an explicit expression for the error you measured in terms of $\mathbf{h}$ and its estimate $\hat{\mathbf{h}}$.