Notes For 10/17/2017

Graph Clustering

Let $G = (V, E)$ be a graph with adjacency matrix A. Our goal is to find clusters in G, i.e. a partition of $V = \{1, ..., n\}$.

Idea: Some permutation of A should have a block structure where the blocks correspond to clusters. We want to find the "hidden blocks," since we don't know the necessary permutation. \

The Stochastic Block Model

For each $i \in \{1, ..., n\}$ we want to find the corresponding cluster label $y_i \in \{1, ..., k\}$, given $\{A_{ij}\}$. Suppose the graph is chosen randomly so that

$$A_{ij} = \begin{cases} 0 & i = j \\ ber(p_{ij}) & i \neq j \end{cases}$$

Since A is symmetric, we must have $A_{ij} = A_{ji}$. The A_{ij} are independent for all $i < j$. The probabilities p_{ij} are given by

$$p_{ij} = \begin{cases} p_1 & y_i = y_j \\ p_2 & y_i \neq y_j \end{cases}$$

for some $0 < p_1 < p_2 < 1$. The goal is to recover $\{y_i\}$ up to some permutation of $\{i\}$.

Special Case: 2 Clusters

$k = 2$. Suppose the clusters have size $\mid c_1 \mid = n_1$, $\mid c_2 \mid = n_2$. Then with some permutation

$$\bar{A} = \mathbb{E}A = \begin{bmatrix} P_1 & P_2 \\ P_2 & P_1 \end{bmatrix}$$

(w/ diagonal 0). Consider $eig(A)$.

$$\bar{A} = \Theta_{kx2} B_{2x2} \Theta_{2xk}^T$$

where $\theta_{il} = \delta_{y_i = l}$ and $B = \begin{bmatrix} p_1 & p_2 \\ p_2 & p_1 \end{bmatrix}$.

So $\bar{A}$ has rank k and the eigenvectors of $\bar{A}$ "indicate" the blocks of A. This can be used to find the clusters.

But we are only given A, not $\bar{A}$. If $eig(A)$ is close to $eig(\bar{A})$, then this method can still work. Write $A = \bar{A} + E$, where $\mathbb{E}E = 0$. We know that $var(E_{ij}) \leq 1$ and that the E_{ij} are independent for $i < j$.

Prop

If $\| \cdot \|_{op}$ denotes the operator norm, then $\| E \| \leq c\sqrt{n}$ for some $c > 0$. If $n_1, n_2 \sim O(n)$, then $\| \bar{A} \|_{op} \sim O(n)$.

Therefore, the eigenvalues of $\bar{A}$ are on a higher order than those of E.

Thm (Davis-Kahan, Stability of Eigenvectors)

Let $\tilde{A} = A + E$, be nxn, symmetric matrices with $\| E \|_{op}$ small. Say that the diagonalization of A is

$$A = U\Lambda U^T = U_1\Lambda_1 U^T + U_2\Lambda_2 U_2^T$$

where $U = [U_1 \mid U_2]$ and U_i is nxn_i, and $\tilde{A} = \tilde{U}\tilde{\Lambda}\tilde{U}^T$. If $\exists$ (a, b) and $\delta > 0$ such that Λ_1 (i.e. diagonal entries of Λ_1)$\sqsubset (a, b)$ and $\Lambda_2 \sqsubset (a - \delta, b + \delta)$ (this is the "spectral gap condition"), then

$$\| U_1^T \tilde{U}_2 \| \leq \frac{\| U_2^T \tilde{E} U_1 \|}{\delta} \leq \frac{\| E \|}{\delta}.$$

Proof

Say $A\psi = \lambda\psi$ and $\tilde{A}\tilde{\psi} = \tilde{\lambda}\tilde{\psi}$. Then $\tilde{\psi}^T A\psi = \lambda(\tilde{\psi}^T\psi)$ and $(A+E)\tilde{\psi} = \tilde{\lambda}\tilde{\psi} \Rightarrow \psi^T A\tilde{\psi} + \psi^T E\tilde{\psi} = \tilde{\lambda}\psi^T\tilde{\psi}$. Combining the last two equations gives $\psi^T E\tilde{\psi} = (\tilde{\lambda} - \lambda)(\psi^T\tilde{\psi})$. Repeating this process for all $\psi, \tilde{\psi}$ and using the spectral gap condition implies the theorem. ///

Notes for 10/19/2017

Topic 4 - Graph Denoising

The idea behind this topic is to use the geometry of a graph to improve estimation of a function on that graph.

Problem: Say $G = (V, E)$,where $V = \{1, ..., n\}$, is a graph with weighted adjacency matrix W, degree matrix D, and $P = D^{-1}W$. We want to estimate the function $f : V \to \mathbb{R}$, or equivalently the vector $f \in \mathbb{R}^n$. $(f_i = f(i))$. We are given the noisy observation $x = f + \epsilon$, where $\epsilon_i \sim N(0, \sigma^2)$. If we knew nothing about G, the "default" estimate would be $f^{MLE} = x$.

Assumption: f is "smooth" on G. First we need to define what this means.

If f were a C^2 function on $[0, 2\pi]$, then we could write it as a Fourier series:

$$f(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx}$$

where e^{ikx} are the eigenfunctions of the laplacian Δ. This means that

$$f^{(n)}(x) = \sum i^n c_k k^n e^{ikx}$$

By Parceval's identity, $\| f^{(n)} \|_{L^2} = \sum \mid c_k k^n \mid$, so for $f^{(n)}$ to exist and have finite L^2 norm, c_k needs to decay faster than $\frac{1}{k^n}$. Therefore, smoothness of f means that the Fourier coefficients decay quickly for large k.

Back to the Graph Setting: We can use this to define smoothness of f.

Prop: Suppose $f = \sum_k c_k \psi_k$, where ψ_k are the eigenvalues of the graph laplacian $L = D - W$, (This is the generalized Fourier series of f.) and that $f^T L f < \delta$. Then, if $1 = \lambda_1 > \cdots > \lambda_k \geq 0$ are the eigenvectors of P,

$$c_k^2 < \frac{\delta}{1-\lambda_k}$$

for $k > 1$.

Remark: $f^T L f = \frac{1}{2} \sum w_{ij}(f_i - f_j)^2$, which is the analogue of $\int_0^{2\pi} f \Delta f dx = \int_0^{2\pi} \mid \nabla f \mid^2 dx$, so $f^T L f$ being small is the analogue of the assumption that the derivative of f is bounded in the continuous case. If $\lambda_k \searrow 0$ fast, then the c_k must also decay fast, so that "f is smooth".

Proof: $f = \sum c_k \psi_k$ and $P\psi_k = \lambda_k \psi_k$. $L = D - W = D(I - P)$, so $f^T L f = f^T D(I-P) f$.

$(I - P) = \sum c_k (1 - \lambda_k) \psi_k$ by the Fourier expansion of f.

$\Rightarrow f^T D(I-P) f = (\sum_l c_l \psi_l)^T D(\sum_k c_k (1-\lambda_k)\psi_k)$. By orthonormality of ψ, this implies $f^T L f = \sum_k d_k c_k^2 (1 - \lambda_k)$, which implies the desired result. ///

Now, the proposed method of estimating f is

$$\hat{f} = Px.$$

Suppose $\epsilon = \sum_k g_k \psi_k$. Note that ϵ, as a noise term, is generally not smooth, so g_k will *not* decay fast. In vector form, $\epsilon = \Psi g$, and $g \sim N(o, \sigma^2 I)$. Suppose G is regular, i.e. $D = v_0 I$, so that $g_k \sim N(0, v_0 \sigma^2)$. The signal-to-noise ratio is $SNR = \frac{|c_k|^2}{\mathbb{E}|g_k|^2} = \frac{|c_k|^2}{v_0 \sigma^2}$.

Now, $\hat{f} = P(f + \epsilon) = \sum (c_k + g_k) \lambda_k \psi_k$ and $f = \sum c_k \psi_k$. From this we can derive

$$(bias)^2(\hat{f}) = \sum \frac{1}{v_0}(1-\lambda_k)^2 c_k^2 \leq \frac{\delta}{v_0}$$

$$var(\hat{f}) = \sigma^2 \sum \lambda_k^2.$$

If f is smooth, then the λ_k decay fast, and $\sigma^2 \sum \lambda_k^2 << \sigma^2 n = var(f^{MLE} = x)$, so $\hat{f}$ has a bias-variance trade-off if f is smooth.

This gives us the motivation for the following estimation method.

The Method of Nonlocal Means

Given data $\{x_i\}_{i=1}^n$, with $x_i \in \mathbb{R}^D$, construct $w_{ij} = exp(- \parallel x_i - x_j \parallel^2 / \epsilon)$. Then let

$$\hat{x}_i = \frac{\sum w_{ij} x_j}{\sum w_{ij}}.$$

Here, the first coordinate $\hat{x}_i(1)$ corresponds to the function $f(i) = \hat{x}_i(1)$ in the analysis above.