EE378B Inference, Estimation and Information Processing

Final: Biclustering

Andrea Montanari

This final is assigned Friday, March 19 at 12:00 pm PST and be due on Saturday, March 20 at 12:00 PM PST. Solutions should be submitted via Gradescope. Please send your solution by email to Andrea (montanari@stanford.edu) at the same time.

You are allowed to use books, notes and papers, but you should provide arguments for anything that was not proved in class. You can also contact Andrea and Qijia for reasonable questions about the text. You are not allowed to consult/collaborate with your colleagues or anybody else.

We will consider a simple model for biclustering. We are given a data matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ which is generated as follows. Partition the set of rows in k equal-sized groups $[m] = R_1 \cup \cdots \cup R_k$, $|R_i| = m/k$, and the set of columns into ℓ equal-sized groups $[n] = C_1 \cup \cdots \cup C_\ell$, $|C_i| = n/\ell$. For a given row index $j \in [m]$, we denote by $\tau_j \in [k]$ the group to which j belongs, and for a given column index $i \in [n]$, we denote by $\sigma_i \in [\ell]$ the index of the group to which i belongs. In other words, $\tau_j = t$ if and only if $j \in R_t$, and $\sigma_i = s$ if and only if $i \in C_s$.
[Throughout we use the notation $[n] := \{1, 2, \dots, n\}$.]

We will write $\boldsymbol{\sigma} = (\sigma_i)_{i\le n}$, $\boldsymbol{\tau} = (\tau_j)_{j\le m}$, for the vertex labels. We will assume that rows and columns are reordered uniformly at random. Equivalently $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are uniformly random vectors with entries $\sigma_i \in [\ell]$, $\tau_j \in [k]$, subject to having equal number of entries of each value.

The distribution of $\boldsymbol{X}$ is parametrized by m, n, k, ℓ, and $v \in \mathbb{R}_{\ge 0}$, and by a matrix $\boldsymbol{\mu} \in \mathbb{R}^{k\times\ell}$. We then set

$$X_{ji} = \mu_{\tau_j,\sigma_i} + v_n\, G_{ji}\,. \tag{1}$$

We will assume for simplicity $(G_{ji})_{j\le m, i\le n} \sim_{iid} \mathsf{N}(0,1)$, and m to be proportional to n, i.e., $m = \delta\cdot n$, for $\delta \in (0,\infty)$. (If you prefer, you can focus on the case $\delta = 1$.) We will assume that $\boldsymbol{\mu}$ is fixed independent of n, and v_n can change with n.

Our objective is to estimate the column labels $\boldsymbol{\sigma} \in [\ell]^n$, up to a global permutation of the label values $\{1, 2, \dots, \ell\}$.

In each question below, you are required to justify your answer. We do not require a fully rigorous proof, but higher credit will be given to more complete and rigorous justifications.

(1) Define a metric for the estimation of $\boldsymbol{\sigma}$, which is invariant under permutation of the label values. Namely, given a permutation $\pi : [\ell] \to [\ell]$, let $\boldsymbol{\sigma}^\pi := (\pi(\sigma_1), \dots, \pi(\sigma_n))$ be the vector obtained by permuting the label values according to π. We want a metric that takes $\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}} \in [\ell]^n$ (where $\boldsymbol{\sigma}$ are the true values, and $\hat{\boldsymbol{\sigma}}$ are the estimates), and returns $Q(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}})$, such that $Q(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}) = Q(\boldsymbol{\sigma}^\pi, \hat{\boldsymbol{\sigma}})$ for all π. Further, we would like this metric to be normalized so that $Q(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}) \in [0, 1]$, with $Q(\boldsymbol{\sigma}, \boldsymbol{\sigma}) = 1$.

(2) A first algorithmic idea would be to sum up each column of $\boldsymbol{X}$, i.e., compute $\boldsymbol{y} = \boldsymbol{X}^{\mathsf{T}}\mathbf{1}$, and then cluster the entries of $\boldsymbol{y}$ to estimate $\boldsymbol{\sigma}$ (this can be done by ordering these entries). Assume $v_n = n^\gamma$. For which values of γ do you expect this method to be successful? For which matrices $\boldsymbol{\mu}$ this method can be successful? Are there matrices $\boldsymbol{\mu}$ for which this approach is never successful?

[Here 'successful' means that the the expected value of the metric defined above converges to 1, $\mathbb{E}Q(\boldsymbol{\sigma}, \hat{\boldsymbol{\sigma}}) \to 1$ as $n \to \infty$.]

(3) Hereafter, assume $\boldsymbol{\mu}\mathbf{1}_\ell = d_1 \cdot \mathbf{1}_k$, and $\boldsymbol{\mu}^{\mathsf{T}}\mathbf{1}_k = d_2 \cdot \mathbf{1}_\ell$ for some constants d_1, d_2. Design a spectral algorithm to estimate $\boldsymbol{\sigma}$. The algorithm should output $\hat{\boldsymbol{\sigma}} \in [\ell]^n$. Again, assume $v_n = n^\gamma$. For which values of γ do you expect your approach to be successful, in the sense of performing better than random guessing in the large n limit?

[Hint: We expect the approach to work well if $\gamma < \gamma_*$, and not for $\gamma > \gamma_*$. You are required to determine γ_*.]

(4) Consider the special case $k = \ell = 2$, and $v_n = \rho n^{\gamma_*}$ with ρ a sufficiently small constant. Will the spectral method at the previous point perform better than random guessing in the large n limit?

(5) As at the previous point, consider the special case $k = \ell = 2$, and $v_n = \rho n^{\gamma_*}$. Implement your spectral method and carry out numerical simulations to confirm your answer at the previous point. How small ρ must be empirically for the algorithm to behave better than random guessing?

(6) Again, consider the special case $k = \ell = 2$. In this case we can rename the labels so that $\boldsymbol{\sigma} \in \{+1, -1\}^n$, and $\boldsymbol{\tau} \in \{+1, -1\}^m$. Consider the following leave-one-out approach. In order to estimate whether $\sigma_i = \sigma_j$:

(*i*) Remove columns i, j from the matrix to construct $\boldsymbol{X}^{(ij)} \in \mathbb{R}^{m \times (n-2)}$.

(*ii*) Apply the spectral method at the previous point to $\boldsymbol{X}^{(ij)}$ to estimate $\boldsymbol{\tau}$, and let $\hat{\boldsymbol{\tau}} \in \{+1, -1\}^m$ be the estimate. If needed, modify this estimate so that $\langle \hat{\boldsymbol{\tau}}, \mathbf{1}_m \rangle = 0$ (assume m even.)

(*iii*) Estimate σ_i, σ_j as $\hat{\sigma}_i = \mathrm{sign}(\langle \hat{\boldsymbol{\tau}}, \boldsymbol{X}\boldsymbol{e}_i \rangle)$, $\hat{\sigma}_j = \mathrm{sign}(\langle \hat{\boldsymbol{\tau}}, \boldsymbol{X}\boldsymbol{e}_j \rangle)$.

As before, set $v_n = \rho\, n^{\gamma_*}$. Considering $n \to \infty$ first, how does $\mathbb{P}(\hat{\sigma}_i\hat{\sigma}_j \neq \sigma_i\sigma_j)$ behave for small ρ? Does it approach 0 as ρ decreases? If yes, how quickly does it approach 0?

[Hint: In order to address these questions, you could try to derive an upper bound of the form $\mathbb{P}(\hat{\sigma}_i\hat{\sigma}_j \neq \sigma_i\sigma_j) \leq F(\rho) + o_n(1)$, and then show that $F(\rho) \to 0$ as $\rho \to 0$.]

(7) Describe a semidefinite programming relaxation of the biclustering problem.