Lecture 16

Last time

- Two-sided bounds on the singular values of sub-Gaussian matrices

Today

- Covariance Estimation
- Clustering a Gaussian Mixture.

Covariance Estimation

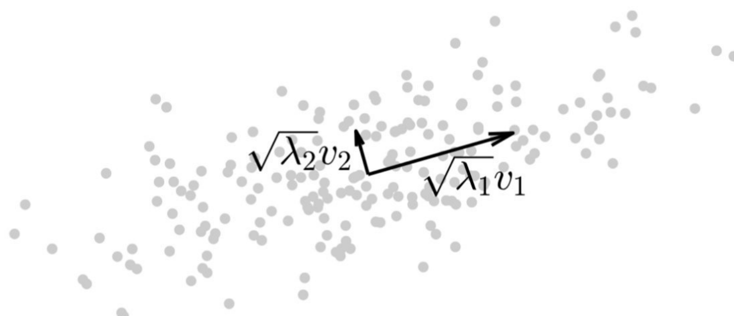
Principal Component Analysis (PCA) is a popular tecnique to reduce the dimension (adaptively). Suppose we have an iid sample $X_1, \dots, X_n \in \mathbb{R}^d$ with $X_i \sim \mathcal{D}$. When d is large, it makes sense to ~~find~~ try to find a projection $\mathcal{Q} : \mathbb{R}^d \to U \subseteq \mathbb{R}^d$ onto a subspace U that encodes "most interesting" dimensions of the distribution $\mathcal{D}$.

Lemma: Suppose $X \sim \mathcal{D}$ whose covariance Σ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$ and eigenvectors $u_1, \dots, u_d$, then $\forall k \in [d]$

$$\lambda_k = \max_{v \perp \{u_1, \dots, u_{k-1}\}, \|v\| = 1} \operatorname{Var}(\langle X, v \rangle)$$

and the maximum is attained at u_k. $\dashv$

Thus, we believe that we can measure how interesting a direction is with its variance, then it makes sense to try to compute the top k for some k eigenvectors of Σ, call them $U_k = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_k \\ | & & | \end{bmatrix}$ and define $Q = U U_k^T$.



200 random points, top eigenvectors scaled by standard deviations.

Issue: We don't have access to Σ. But, we can approximate it using samples via

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T.$$

Because of the law of large numbers

we know that $\Sigma_n \to \Sigma$ a.s. But how large does n have to be for $\|\Sigma_n - \Sigma\|_{op} \leq \varepsilon$ w.h.p?

Theorem: Let X be a sub-Gaussian random vector in $\mathbb{R}^d$ with $\mathbb{E}X = 0$ and $\mathbb{E}XX^T = \Sigma$. Moreover, assume

$K \geq 1$

$$\|\langle X, x\rangle\|_{\psi_2} \leq K\, x^T \Sigma x \quad \forall x \in \mathbb{R}^d.$$

Then, for all m, we have

$$\frac{\mathbb{E}\|\Sigma_n - \Sigma\|_{op}}{\|\Sigma\|_{op}} \leq CK^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right).$$

Proof: To apply the main result from lecture 15, we modify X_i to make them isotropic. In particular we let

$$Z = \Sigma^{-1/2} X \quad \text{and} \quad Z_i = \Sigma^{-1/2} X_i.$$

Then, it is not hard to check that

$$\mathbb{E} Z = 0, \quad \mathbb{E} ZZ^T = I, \quad \|Z\|_{\psi_2} \leq K.$$

Hence, we can rewrite

$$\|\Sigma_n - \Sigma\|_{op} = \left\|\Sigma^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T - I\right)\Sigma^{1/2}\right\|_{op}$$

For diagonal matrices $\|D^{1/2}\|_{op}^2 = \|D\|_{op}$

$$\le \|\Sigma\|_{op} \Big\| \underbrace{\frac{1}{n}\sum_{i=1}^{n} z_i z_i^T - I}_{R_n} \Big\|_{op}. \quad (♡)$$

Thus, if we consider the matrix A with rows given by z_i^T, we get

$$A^T A = \sum_{i=1}^{n} z_i z_i^T.$$

Applying $(\overset{\wedge\wedge}{\omega})$ from lecture 15 gives

$$\|R_n\|_{op} \le CK^2\left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)$$

substituting this into (♡) completes the proof. $\square$

Corollary: Consider the setting of the previous theorem. There $\exists C > 0$ s.t. for all $\varepsilon \in (0,1)$ if

$$n \ge \frac{C\varepsilon^{-2} n}{K^4}$$

Then,

$$\mathbb{E}\|\Sigma_n - \Sigma\|_{op} \le \varepsilon \|\Sigma\|_{op}. \qquad \dashv$$

Clustering Gaussian Mixtures

Let's illustrate another type of

clustering application, this time for point clouds as opposed to networks.

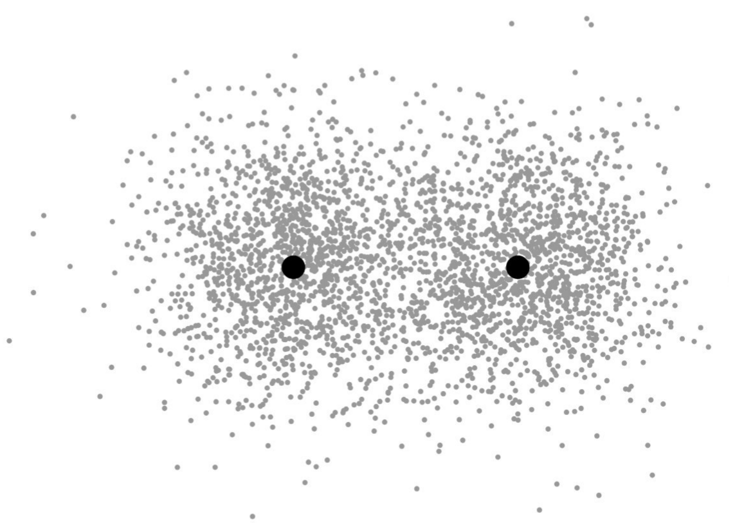
Def (Gaussian Mixture Model): Generate n random points in $\mathbb{R}^n$ iid as follows:

1. Flip a fair coin s_i.
2. Draw a point X from $N(s_i\mu, I_d)$.

Analogously, we could define

$$X = \underbrace{s}_{\sim \mathrm{Unif}\{\pm 1\}}\mu + \underbrace{g}_{\sim N(0, I_d)}.$$

with s and g independent.



$n = 3000$ points drawn from GMM with means $\pm\mu = \pm(1.6, 0)$.

Given observations $X_1, \dots, X_n$ our goal is to estimate the labels $s_1, \dots, s_n$. Once more we use a spectral method that tries to look for a direction of maximum variance.

Spectral Clustering Algorithm

Input: Samples $X_1, \dots, X_n \in \mathbb{R}^d$

1. Compute the sample covariace
$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T.$$
2. Compute the top eigenvector
$$v = u_1(\Sigma_n).$$
3. For all $i \in [n]$, output
$$\hat{s}_i = \text{sign}(\langle X, v \rangle).$$

What we have proven so far can be used to establish the following result.

Theorem: Let $X_1, \dots, X_n$ be points in $\mathbb{R}^d$ drawn from $GMM(\mu)$. There exists $C>0$ s.t. if $n \geq Cd$ and $\|\mu\|_2 \geq C_1$ then with probability at least 0.99 the Spectral Clustering Algorithm only misclassifies at most 1% of the points. $\dashv$

Prove this result!

Remark: 1) The diameter of a point cloud drawn from $N(0, I_d)$ is on the order of $\sqrt{n}$, yet a small amount of separation $\|\mu\| \asymp 1$ suffices for classification.

2) This is optimal (one cannot do better than $n \geq Cd$. However when $g \sim N(0, \Sigma)$ with Σ unknown, the picture becomes way more nuanced. See "Clustering a mixture of Gaussian with unknown covariance" by D. Davis, K. Wang & the instructor.