36-755: Advanced Statistical Theory 1 **Fall 2016**

Lecture 13: October 17

Lecturer: Alessandro Rinaldo *Scribes: Adarsh Prasad*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

13.1 Sparse PCA for Spiked Covariances

Setup. Let $\Sigma_{d\times d} = \theta vv^T + I_d$ where $\theta > 0$, $v \in \mathcal{S}^{d-1}$ and $v \in \mathbb{R}$ s.t. $||v||_0 = k \leq \frac{d}{2}$. Observe that $\|\Sigma\|_{op} = 1 + \theta$. Then the goal is to estimate v by solving the following optimization problem:

$$\widehat{v} \in \underset{\substack{u\in\mathcal{S}^{d-1} \\ ||u||_0 = k' \\ k\leq k'\leq \frac{d}{2}}}{\operatorname{argmax}} u^T\widehat{\Sigma}u$$

For the above setup, we have the following Theorem:

Theorem 13.1 *Let $\{X_1, X_2, \ldots, X_n\}$ be zero-mean with co-variance Σ, and each $X_i \in SG_d(||\Sigma||_{op})$, then w.p. atleast $1-\delta$, $\delta \in (0,1)$:*

$$\min_{\epsilon\in\pm 1} ||\epsilon\widehat{v} - v||^2 \precsim \frac{1+\theta}{\theta} \max\{\sqrt{A}, A\}$$
$$\textit{where } A = \frac{(k+k')\log\left(\frac{ed}{k+k'}\right) + \log(1/\delta)}{n}$$

Proof:(Theorem 13.1)
Observe the following:

$$\begin{aligned}\theta \sin^2\left(\angle(v,\widehat{v})\right) &= v^T\Sigma v - \widehat{v}^T\Sigma v \\ &\leq \left\langle\left\langle \widehat{\Sigma} - \Sigma, \widehat{v}\widehat{v}^T - vv^T \right\rangle\right\rangle \end{aligned} \tag{13.1}$$

where $\langle\langle A, B\rangle\rangle = \text{trace}(A^TB)$. Also, observe that the frobenius norm $||A||_F^2 = \langle\langle A, A\rangle\rangle$.

Now, we know that $v, \widehat{v}$ are k, k'-sparse respectively, which implies that $\exists S \subset \{1, 2, \ldots, d\}$ with $|S| \leq k + k'$, such that:

$$\left\langle\left\langle \widehat{\Sigma} - \Sigma, \widehat{v}\widehat{v}^T - vv^T \right\rangle\right\rangle = \left\langle\left\langle \widehat{\Sigma}_S - \Sigma_S, \widehat{v}_S\widehat{v}_S^T - v_Sv_S^T \right\rangle\right\rangle$$

where $\widehat{\Sigma}_S, \Sigma_S$ are sub-matrices of $\widehat{\Sigma}, \Sigma$ respectively with rows/columns in S. Similarly, $\widehat{v}_S, v_s$ are sub-vectors of $\widehat{v}, v$ respectively with entries in S.

Now, we get:

$$\left\langle\left\langle \widehat{\Sigma}-\Sigma, \widehat{v}\widehat{v}^T - vv^T\right\rangle\right\rangle = \left\langle\left\langle \widehat{\Sigma}_S-\Sigma_S, \widehat{v}_S\widehat{v}_S^T - v_Sv_S^T\right\rangle\right\rangle$$
$$\leq \left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \left\|\widehat{v}_S\widehat{v}_S^T - v_Sv_S^T\right\|_1$$
$$\left\langle\left\langle \widehat{\Sigma}-\Sigma, \widehat{v}\widehat{v}^T - vv^T\right\rangle\right\rangle \leq \left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \times \sqrt{2}\left\|\widehat{v}_S\widehat{v}_S^T - v_Sv_S^T\right\|_2 \tag{13.2}$$

Now, we have that,

$$\left\|\widehat{v}_S\widehat{v}_S^T - v_Sv_S^T\right\|_2 = \sqrt{2(1-(\widehat{v}^Tv)^2)} \quad \textit{(proved in HW-5)}$$
$$= \sqrt{2\sin^2\left(\angle(v,\widehat{v})\right)} \tag{13.3}$$

Plugging the above in Equation 13.2, we get:

$$\left\langle\left\langle \widehat{\Sigma}-\Sigma, \widehat{v}\widehat{v}^T - vv^T\right\rangle\right\rangle \leq \left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \times 2 \times \sin\left(\angle(v,\widehat{v})\right) \tag{13.4}$$

Combining Equation 13.1 and Equation 13.4, we get the following result:

$$\theta \sin\left(\angle(v,\widehat{v})\right) \leq 2\left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \tag{13.5}$$

Now, recall that $\min_{\epsilon\in\pm1} ||\epsilon\widehat{v}-v||^2 \leq 2\sin^2\left(\angle(v,\widehat{v})\right)$. Plugging this into Equation 13.5, we get:

$$\min_{\epsilon\in\pm1} ||\epsilon\widehat{v}-v|| \leq \frac{\sqrt{8}}{\theta}\left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \leq \frac{\sqrt{8}}{\theta}\sup_{S:|S|=k+k'}\left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \tag{13.6}$$

Now, to control the probability of deviation of supremum, we use union bound:

$$P\left(\sup_{S:|S|=k+k'}\left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \geq t\left\|\Sigma\right\|_{op}\right) \leq \binom{d}{k+k'} \times 144^{k+k'} \times \exp\left(-n/2.\min\left[\left(\frac{t}{32}\right)^2, \frac{t}{32}\right]\right) \tag{13.7}$$

Using bounds on binomial coefficients: $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$, Plugging this into Equation 13.7 and moving everything into the exponential, we get:

$$P\left(\sup_{S:|S|=k+k'}\left\|\widehat{\Sigma}_S-\Sigma_S\right\|_{op} \geq t\left\|\Sigma\right\|_{op}\right) \leq \exp\left(-\frac{n}{2}\min\left[\left(\frac{t}{32}\right)^2, \frac{t}{32}\right] + 2(k+k')\log 12 + (k+k')\log\left(\frac{ed}{k+k'}\right)\right) \tag{13.8}$$

Now, to prove Theorem 13.1, pick $t \geq \max\{A, \sqrt{A}\}$ such that RHS of Equation 13.8 is less than equal to δ, so, choosing $A \precsim \frac{(k+k')\log\left(\frac{ed}{k+k'}\right)+\log(1/\delta)}{n}$ is sufficient.
For this value of A, we get that with probability atleast $1-\delta$:

$$\min_{\epsilon\in\pm1} ||\epsilon\widehat{v}-v||^2 \precsim \frac{\|\Sigma\|_{op}}{\theta}\max\{\sqrt{A}, A\}$$

■

Remark!!: For proof for a broader class of covariances, refer to Theorem 8.1 in [1]

13.2 Uniform Law of Large Numbers (Chapter 4 [1])

Consider the following Example:
Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P$ with CDF F. i.e. $F(x) = P(X \leq x)$, $\forall x \in \mathbb{R}$. Now, for a fixed t, we want to estimate the empirical CDF $\widehat{F}_n(t)$ given by:

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq t)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

$\widehat{F}_n(t)$ concentrates well around $F(t)$. To see this, observe that $\mathbb{I}(X_i \leq t) \sim \text{Bernoulli}(F(t))$, so one can use Hoeffding's inequality to get tight concentration. This means that one can estimate F very well, ***point-wise.***

However, a better(stronger) result is to bound $\left\|\widehat{F}_n - F\right\|_\infty = \sup_{z \in \mathbb{R}} |\widehat{F}_n(z) - F(z)|$. One would want tight bounds on $P\left(\left\|\widehat{F}_n - F\right\|_\infty \geq t\right)$.

General Setup. Let P be a probability distribution on $(\mathscr{X}, \mathscr{B})$ and $\mathscr{F}$ be a class of real valued functions on $\mathscr{X}$. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P$ and construct the empirical measure associated to the sample:

$$\forall A \text{ measurable}, \quad P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \in A)$$

We make **no** assumptions on $\mathscr{F}$ apart from that it is a uniformly bounded class. We are interested in the random variable:

$$\|P_n - P\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \left(\frac{1}{n} \left| \sum_{i=1}^{n} (f(X_i) - E[f(X_i)]) \right| \right) = \sup_{f \in \mathscr{F}} \|P_n f - Pf\|$$

We want to establish convergence in probability, i.e. $\|P_n - P\|_{\mathscr{F}} \overset{P}{\to} 0$. If we can establish this convergence, then $\mathscr{F}$ is called a **Glivenko-Cantelli** class.

Definition 13.2 *We say that $\mathscr{F}$ is a Glivenko-Cantelli class for P if $\|P_n - P\|_{\mathscr{F}}$ converges to zero in probability as $n \to \infty$.*

In the following example, we show how the uniform concentration of the empirical CDF is just a special case of the above definition.

Example 13.3 (Glivenko-Cantelli Theorem) *For any distribution, the empirical CDF F_n is a strongly consistent estimator of the population CDF F in the uniform norm, meaning that:*

$$\left\|\widehat{F}_n - F\right\|_\infty \overset{a.s.}{\to} 0$$

Consider the function class $\mathscr{F} = \left(\mathbb{I}_{(-\infty,t]}(\cdot) | t \in \mathbb{R}\right)$ where $\mathbb{I}_{(-\infty,t]}(\cdot)$ is $\{0-1\}$ valued indicator function of the interval $(-\infty, t]$. For each fixed $t \in \mathbb{R}$, we have $E[\mathbb{I}_{(\infty,t]}(X)] = P[Xt] = F(t)$, so that the classical Glivenko-Cantelli theorem corresponds to a strong uniform law for the class in Definition 13.2.

13.2.1 Decision-Theoretic Motivation.

Consider an indexed-family of probability distributions $\mathscr{P} = \{P_\theta | \theta \in \Omega \subseteq \Theta\}$, and suppose that we are given n samples $X^n = \{X_1, \ldots, X_n\}$ each sample lying in some space $\mathscr{X}$ and suppose that the samples are drawn i.i.d. according to a distribution P_{θ^*} for some fixed but unknown $\theta^* \in \Theta$.

A standard decision-theoretic approach to estimating θ^* is based on minimizing a loss function of the form $\mathcal{L}_\theta(x)$, which measures the *discrepancy* between a parameter $\theta \in \Omega$ and the sample $x \in \mathscr{X}$.

Given the collection of n samples X^n, the *empirical risk*($\widehat{\mathcal{R}}_n(\theta, \theta^*)$) associated to $\mathcal{L}_\theta(\cdot)$ is defined by:

$$\widehat{\mathcal{R}}_n(\theta, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\theta(X_i)$$

The *risk*(or *population risk*) is defined by:

$$\mathcal{R}(\theta, \theta^*) = E_{\theta^*} \left[\mathcal{L}_\theta(X)\right]$$

where the expectation E_{θ^*} is taken over a sample $X \sim P_{\theta^*}$.

Let $\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \Omega} \widehat{\mathcal{R}}_n(\theta, \theta^*)$ be the empirical risk minimizer, then one is interested in the *excess-risk* $\delta\mathcal{R}(\widehat{\theta}, \theta^*)$ defined by:

$$\delta\mathcal{R}(\widehat{\theta}, \theta^*) = \mathcal{R}(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega} \mathcal{R}(\theta, \theta^*)$$

Example 13.4 (Maximum Likelihood) *Consider a family of distributions $\{P_\theta, \theta \in \Theta\}$, each with a strictly positive density p(defined with respect to a common underlying measure). In order to estimate the true parameter, consider the loss function given by:*

$$\mathcal{L}_\theta(x) = \log\left(\frac{P_{\theta^*}(x)}{P_\theta(x)}\right)$$

Then the population risk $\mathcal{R}(\theta, \theta^*)$ is simply the KL-divergence $KL(P_{\theta^*}||P_\theta)$. The empirical risk minimizer $\widehat{\theta}$ is the Maximum Likelihood Estimator. To see this:

$$\begin{aligned}
\widehat{\theta} &\in \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\theta(X_i) \\
&\in \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{P_{\theta^*}(x)}{P_\theta(X_i)}\right) \\
&\in \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{1}{P_\theta(X_i)}\right) \\
&\in \operatorname*{argmax}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log\left(P_\theta(X_i)\right) \\
&\in \operatorname*{argmax}_{\theta} P_\theta(X^n)
\end{aligned}$$

Example 13.5 (Binary Classification) *Suppose that we are given n samples $(X_i, Y_i) \in \mathbb{R}^d \times \{1, +1\}$, where X_i corresponds to a set features, and the binary variable Y_i corresponds to a label. This data can be viewed as being generated from a distribution P_X over the features and a (binary-valued)conditional distribution $P_{Y|X}$, then define the likelihood ratio $\psi(x) = \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$.*

Then the goal of binary classification is to estimate a function $f : \mathbb{R}^d \to \{-1, 1\}$ such that probability of mis-classification $P(Y \neq f(X)$ is minimized.

Consider the loss function:

$$\mathcal{L}_f^{0/1}(X, Y) = \begin{cases} 1 & \text{if } Y \neq f(X) \\ 0 & \text{otherwise.} \end{cases}$$

Observe that the population risk for zero-one loss function $\mathcal{L}_f^{0/1}$ is the probability of mis-classification $P(Y \neq f(X)$.

The function that minimizes this probability of mis-classification(or 0-1 population risk) is called the Bayes-classifier f^* and in the case of $P(Y = 1) = P(Y = -1) = 1/2$, the Bayes classifier $f^*(x) = sign(\psi(x) - 1/2)$.

$$f^*(X) = \begin{cases} 1 & \text{if } \psi(X) \geq 1/2 \\ -1 & \text{if } \psi(X) < 1/2 \end{cases}$$

For any $f : \mathbb{R}^d \to \{-1, 1\}$, the empirical risk for $\mathcal{L}^{0/1}$ given by:

$$\widehat{\mathcal{R}}_n(f, f^*) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{f(X_i) \neq Y_i\}$$

is the number of training sample mis-classified.

Let $\widehat{\theta}$ be the empirical risk minimizer, then the excess risk $\delta\mathcal{R}(\widehat{\theta}, \theta^*) = \mathcal{R}(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega} \mathcal{R}(\theta, \theta^*)$ can be written as:

$$\begin{aligned} \delta\mathcal{R}(\widehat{\theta}, \theta^*) &= \mathcal{R}(\widehat{\theta}, \theta^*) - \inf_{\theta \in \Omega} \mathcal{R}(\theta, \theta^*) = \mathcal{R}(\widehat{\theta}, \theta^*) - \mathcal{R}(\theta_0, \theta^*) \\ &= \underbrace{\mathcal{R}(\widehat{\theta}, \theta^*) - \widehat{\mathcal{R}}_n(\widehat{\theta}, \theta^*)}_{T_1} + \underbrace{\widehat{\mathcal{R}}_n(\widehat{\theta}, \theta^*) - \widehat{\mathcal{R}}_n(\theta_0, \theta^*)}_{T_2} + \underbrace{\widehat{\mathcal{R}}_n(\theta_0, \theta^*) - \mathcal{R}(\theta_0, \theta^*)}_{T_3} \end{aligned}$$

Note that $T_2 \leq 0$ as $\widehat{\theta}$ is the minimizer of empirical risk over Ω. T_3 can be dealt with in a relatively straightforward manner, because θ_0 is a deterministic(unknown) quantity.

To control T_1, observe that it can be written as:

$$\begin{aligned} T_1 &= \mathcal{R}(\widehat{\theta}, \theta^*) - \widehat{\mathcal{R}}_n(\widehat{\theta}, \theta^*) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(E[\mathcal{L}_{\widehat{\theta}}(X_i)] - \mathcal{L}_{\widehat{\theta}(X_i)} \right) \\ &\leq \sup_{\theta \in \Omega} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathcal{L}_{\widehat{\theta}(X_i)} - E[\mathcal{L}_{\widehat{\theta}}(X_i)] \right) \right| = \|P_n - P\|_{\mathscr{L}(\Omega)} \end{aligned}$$

where $\mathscr{L}(\Omega) = \{\mathcal{L}_\theta; \theta \in \Omega\}$
Note that T_3 is also dominated by $\|P_n - P\|_{\mathscr{L}(\Omega)}$. So, to control the excess risk of empirical risk minimizers, one needs to establish a uniform law of large numbers for the loss class $\mathscr{L}(\Omega)$.

References

[1] M. Wainwright, High-dimensional statistics: A non-asymptotic viewpoint