36-755: Advanced Statistical Theory I **Fall 2017**

Lecture 15: October 23

Lecturer: Alessandro Rinaldo *Scribes: Jingyan Wang*

15.1 Interpretations for Principal Component Analysis

We consider a random vector $X \in \mathbb{R}^d$, with mean $\mathbb{E}[X] = 0$ and covariance $\mathbb{V}[X] = \Sigma$. Let $\Sigma = V \Lambda V^T$, where V is a $d \times d$ orthogonal matrix, and Λ is a $d \times d$ diagonal matrix, with diagonal entries corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$.

15.1.1 Interpretation 1: Direction of Maximal Variance

The first eigenvector is the direction of maximal variance. Namely, the optimal solution to the following maximization problem is v_1:

$$v^* = \operatorname*{argmax}_{v \in S^{d-1}} \operatorname{var}(v^T X) = v_1,$$

with $\operatorname{var}(v_1^T X) = \lambda_1$, and projected data $v_1(v_1^T X)$.

More generally, let $V_{d,r} = \{d \times r \text{ matrices with orthonormal columns}\}$. The projection onto an r-dimensional subspace with the maximal variance,

$$v^* = \operatorname*{argmax}_{V \in V_{d,r}} \mathbb{E} \left\| V^T X \right\|^2$$

is given by the first r eigenvectors of Σ, namely, $[v_1, \cdots, v_r]$, and the maximum is $\sum_{i=1}^r \lambda_i$.

15.1.2 Interpretation 2: Approximation to the Covariance Matrix

Suppose we want to find a $d \times d$ matrix Z^* with rank at most r, such that

$$Z^* = \operatorname*{argmin}_{\operatorname{rank}(Z) \leq r} \|\Sigma - Z\|_F^2.$$

The solution is $Z^* = \sum_{i=1}^r \lambda_i v_i v_i^T$ with $\|Z^* - \Sigma\|_F^2 = \sum_{i=r+1}^d \lambda_i^2$.

This is a notably a non-convex problem that has a close-form solution.

15.1.3 Interpretation 3: Subspace Interpolation

Suppose we want to find a linear subspace S^* of dimension r, such that

$$S^* = \operatorname*{argmin}_S \mathbb{E}\|X - \Pi_S X\|,$$

where Π_S is the orthogonal projection onto S.

S^* is the subspace spanned by the r largest eigenvectors. Namely,

$$\Pi_S = V_r V_r^T, \quad \text{with the } d \times r \text{ matrix } V_r = \begin{bmatrix} v_1 & \cdots & v_r \end{bmatrix}.$$

The corresponding approximation error is $\sum_{i=r+1}^d \lambda_i^2$.

15.2 Perturbation Theory

We consider the problem of estimating the eigenvalues and eigenvectors of Σ, where we can only observe the empirical covariance matrix $\hat{\Sigma} = \Sigma + E$, where the noise matrix E is symmetric.

15.2.1 Estimating eigenvalues

It is easy to see that

$$\begin{aligned} \lambda_{max}(\hat{\Sigma}) &= \max_{v \in S^{d-1}} v^T(\Sigma + E)v \\ &\leq \lambda_1 + \max_{v \in S^{d-1}} |v^T E v| \\ &= \lambda_1 + \|E\|_{op} \end{aligned}$$

Hence, if we approximate the first eigenvalue λ_1 by $\hat{\lambda}_1 = \lambda_{max}(\hat{\Sigma})$, the error is bounded by $|\hat{\lambda}_1 - \lambda_1| \leq \|E\|_{op}$.

In fact, we have the following theorem:

Theorem 15.1 (Weyl's Theorem) *Let $\hat{\Sigma} = \Sigma + E$, where Σ and E are symmetric matrices. Let λ_i and $\hat{\lambda}_i$ be the i^{th} eigenvalues of Σ and $\hat{\Sigma}$. Then,*

$$\max_{i=1,\cdots,d} \quad |\hat{\lambda}_i - \lambda_i| \leq \|E\|_{op}$$

Remark 15.2 *To estimate eigenvalues of Σ well using $\hat{\Sigma}$, we need $\|E\|_{op} = \|\hat{\Sigma} - \Sigma\|_{op}$ to vanish as $n \to \infty$.*

15.2.2 Estimating eigenvectors

The eigenvalues are stable under perturbation. However, the eigenvectors are not. To see that, consider the following example:

Example 15.3 *Let $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $E = \begin{bmatrix} 0 & \epsilon \\ \epsilon & 0 \end{bmatrix}$. Then, we have*

$$\hat{\Sigma} = \Sigma + E = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}.$$

The eigenvalues of Σ *are* 1 *and* 1*. The eigenvectors of* Σ *are* $\begin{bmatrix}1\\0\end{bmatrix}$ *and* $\begin{bmatrix}0\\1\end{bmatrix}$.

The eigenvalues of $\hat{\Sigma}$ *are* $1+\epsilon$ *and* $1-\epsilon$ *(we have* $\|\hat{\Sigma}-\Sigma\|_{op} = \|E\|_{op} = \epsilon$*). However, for any* ϵ*, the eigenvectors of* $\hat{\Sigma}$ *are* $\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$ *and* $\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$.

From this example, we see that the eigenvectors are in general not stable under perturbation. The issue in this example is that the eigenvalues of Σ are not well-separated (so called "eigengap").

We will introduce bounds for the distance between eigenvectors under perturbation, and more generally, between subspaces spanned by eigenvectors. For that, we need some more definitions.

15.3 Distance between Subspaces

15.3.1 Definitions

Let $\mathcal{E}$ and $\mathcal{F}$ be r-dimensional subspaces of $\mathbb{R}^d$, with $r \leq d$. Let $\Pi_{\mathcal{E}}$ and $\Pi_{\mathcal{F}}$ be projectors onto these two subspaces.

We first consider the angle between two vectors, that is, when $r = 1$. Then, $\mathcal{E}$ and $\mathcal{F}$ are spanned by vectors. Denote the two corresponding vectors as $v_1, v_2 \in S^{d-1}$. The angle between v_1 and $v_2$is defined as:

$$\angle(v_1, v_2) = \cos^{-1}|v_1^T v_2|$$

Now, we need to extend this concept to subspaces (when $r > 1$).

Let E and F be $d \times r$ matrices with orthonormal columns such that range$(E) = \mathcal{E}$ and range$(F) = \mathcal{F}$. Then, the projectors can be written as $\Pi_{\mathcal{E}} = EE^T$ and $\Pi_{\mathcal{F}} = FF^T$.

Now, we define the angle between subspaces $\mathcal{E}$ and $\mathcal{F}$ as the following:

Definition 15.4 *The canonical or principal angles between* $\mathcal{E}$ *and* $\mathcal{F}$ *are:*

$$\theta_1 = \cos^{-1}\sigma_1, \cdots, \theta_r = \cos^{-1}\sigma_r,$$

where $\sigma_1, \cdots, \sigma_r$ *are singular values of* $E^T F$ *or* $F^T E$.

A general result known as CS-decomposition in linear algebra gives the following (see [SS90]):

$$E^T F = U \cos\Theta V^T,$$

where $\Theta = \begin{bmatrix}\theta_1 & \cdots & 0\\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_r\end{bmatrix}$.

An alternative way to define canonical angles is the following:

Definition 15.5 *Let the 1^{st} canonical angle be:*

$$\cos^{-1}(\max_{\substack{x\in\mathcal{E}\\ \|x\|=1}} \max_{\substack{y\in\mathcal{F}\\ \|y\|=1}} \left|x^T y\right|) = \cos^{-1}(\left|x_1^T y_1\right|),$$

where x_1 and y_1 are maximizers.

The k^{th} canonical angle is (for $k = 2, \cdots, r$):

$$\cos^{-1}(\max_{\substack{x\in\mathcal{E},\\ \|x\|=1\\ x^T x_{k'}=0\\ for\ \forall k'=1,\cdots,k-1}} \max_{\substack{y\in\mathcal{F},\\ \|y\|=1\\ y^T y_{k'}=0\\ for\ \forall k'=1,\cdots,k-1}} |x^T y|)$$

Another way of defining canonical angles is the following:

Definition 15.6 *Let $\theta_k = \sin^{-1}(s_k)$ for $k = 1, \cdots, r$, where $s_1, \cdots, s_k$ are the singular values of*

$$\Pi_{\mathcal{E}}(1 - \Pi_{\mathcal{F}}) = EE^T(I - FF^T) = U \sin \Theta V^T.$$

Now, given the definition of the canonical angles, we can define the distances between subspaces $\mathcal{E}$ and $\mathcal{F}$ as the following:

Definition 15.7 *The distance between $\mathcal{E}$ and $\mathcal{F}$ is $\|\sin\Theta\|_F$, which is a metric over the space of r-dimensional linear subspaces of $\mathbb{R}^d$.*

Equivalently,

$$\begin{aligned}\|\sin\Theta(\mathcal{E},\mathcal{F})\|_F^2 =& \|\Pi_{\mathcal{E}}(I - \Pi_{\mathcal{F}})\|_F^2\\ =& \|\Pi_{\mathcal{F}}(I - \Pi_{\mathcal{E}}\|_F^2\\ =& \frac{1}{2}\|\Pi_{\mathcal{E}} - \Pi_{\mathcal{F}}\|_F^2.\end{aligned}$$

15.3.2 Davis-Kahan Theorem

The Davis-Kahan Theorem provides a bound on subpaces spanned by eigenvectors under perturbation.

Theorem 15.8 (Davis-Kahan Theorem) *(Thm V.3.6 in [SS90]) Let Σ and $\hat{\Sigma}$ be $d \times d$ symmetric matrices with eigenvalues*

$$\begin{aligned}\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d\\ \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_d\end{aligned}$$

Fix $1 \leq r \leq s \leq d$, and let V and $\hat{V}$ be $d \times (s - r + 1)$ matrices with orthonormal columns corresponding to eigenvalues $\{\lambda_j\}_{j=r}^s$ and $\{\hat{\lambda}_j\}_{j=r}^s$. Let $\mathcal{E}$ and $\mathcal{F}$ be the subspaces spanned by columns of V and $\hat{V}$.

Define the eigengap as

$$\delta = \inf\{\left|\lambda - \hat{\lambda}\right| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s+1}] \cup [\hat{\lambda}_{r-1}, \infty)\},$$

where we define $\hat{\lambda}_0 = -\infty$ *and* $\hat{\lambda}_{d+1} = \infty$. *See Figure 15.1 for an illustration.*

If $\delta > 0$, *then*

$$\|\sin\Theta(\mathcal{E}, \mathcal{F})\|_F \leq \frac{\|\hat{\Sigma} - \Sigma\|_F}{\delta}$$

The result also holds for the operator norm $\|\cdot\|_{op}$.

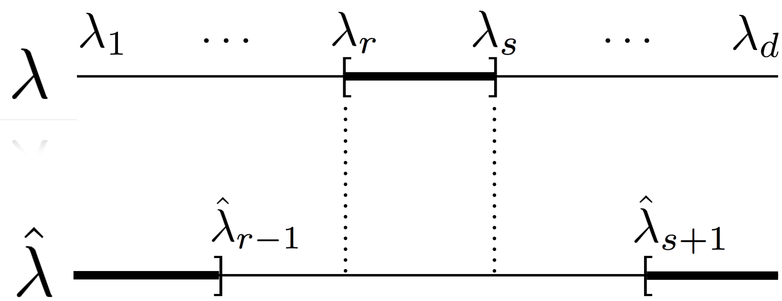

Figure 15.1: Illustration of the eigengap δ.

Remark 15.9 *In practice, we may assume* $\|\hat{\Sigma} - \Sigma\|_{op} \leq \gamma_n$, *and*

$$\left|\hat{\lambda}_{s+1} - \lambda_s\right| \geq \lambda_s - \lambda_{s+1} - \gamma_n$$
$$\left|\hat{\lambda}_{r-1} - \lambda_r\right| \geq \lambda_{r-1} - \lambda_r - \gamma_n.$$

If $\delta^* = \min\{\lambda_s - \lambda_{s+1}, \lambda_{r-1} - \lambda_r\} \geq \gamma_n$, *then, the bound implied by Davis-Kahan is:*

$$\frac{\|\hat{\Sigma} - \Sigma\|_F}{\delta^* - \gamma_n}.$$

If $r = 1$, *then* $\delta^* = \lambda_s - \lambda_{s+1}$.

If γ_n *is small enough such that* $\delta^* - \gamma_n \geq \frac{\lambda_s - \lambda_{s-1}}{2}$, *the bound becomes* $\frac{2\|\hat{\Sigma} - \Sigma\|_F}{\delta^*}$.

Next lecture, we will introduce a variant of the Davis-Kahan Theorem.

References

[SS90] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory.* Academic Press, 1990.