STAT538A: Statistical Learning: the Way of the Kernel Winter 2020

Lecture 6

Lecturer: Zaid Harchaoui *Scribe: Lang Liu*

1 Kernel principal component analysis

1st kernel principal component. From previous lecture we know that, for Kernel PCA, the 1st kernel principal component $f_1 \in \mathcal{H}$ admits a representation of the form

$$f_1 = \sum_{i=1}^n \alpha_i^{(1)} k(x_i, \cdot),$$

where $\alpha^{(1)} = (\alpha_1^{(1)}, \dots, \alpha_n^{(1)})^\top$ can be obtain via

$$\alpha^{(1)} = \operatorname*{arg\,max}_{\alpha \in^n} \frac{\alpha^\top K^2 \alpha}{\alpha^\top K \alpha}.$$

***i*-th kernel principal component.** Similar to KPC1, the i-the kernel principal component is given by

$$f_i = \operatorname*{arg\,max}_{f \perp \operatorname{span}\{f_1, \cdots, f_{i-1}\}} \frac{1}{(n-1)\|f\|_{\mathcal{H}}^2} \sum_{j=1}^n f(x_j)^2.$$

It can also be transformed into finding the corresponding parameter $\alpha^{(i)}$:

$$\begin{aligned} \alpha^{(i)} = \operatorname*{arg\,max}_{\alpha \in \mathbb{R}^n} & \ \alpha^\top K^2 \alpha \\ s.t. & \ \alpha^\top K \alpha = 1, \alpha_j^\top K \alpha = 0, \forall j < i. \end{aligned}$$

Solution to α. Analogous to PCA, we can find KPCs through eigenvalue decomposition of K. Assume $K \in \mathbb{R}^{n \times n}$ is non-singular, and let $\beta = K^{1/2}\alpha$ so that $\alpha = K^{-1/2}\beta$. Now the problem becomes

$$\begin{aligned} \beta^{(i)} = \operatorname*{arg\,max}_{\beta \in \mathbb{R}^n} & \ \beta^\top K \beta \\ s.t. & \ \beta^\top \beta = 1, \beta_j^\top \beta = 0, \forall j < i. \end{aligned}$$

This is exactly an eigenvalue problem and the solution is

$$\beta^{(i)} = \text{the } i\text{-th leading eigenvector of } K,$$

and thus $\alpha^{(i)} = K^{-1/2}\beta^{(i)} = \beta^{(i)}/\sqrt{\lambda_i}$, where λ_i is the corresponding eigenvalue of $\beta^{(i)}$. In fact, assume that K admits an eigenvalue decomposition $K = U\Lambda U^\top$ in which U is an orthogonal matrix, then $\beta^{(i)}$ is just the i-th column of U. Hence,

$$\alpha^{(i)} = K^{-1/2}\beta^{(i)} = U\Lambda^{-1/2}U^\top \beta^{(i)} = \frac{u_i}{\sqrt{\lambda_i}} = \frac{\beta^{(i)}}{\sqrt{\lambda_i}}.$$

Algorithm 2 Kernel PCA

```
Center the kernel matrix
    $K \leftarrow (I_n - \frac{\mathbf{1}\mathbf{1}^\top}{n})K(I_n - \frac{\mathbf{1}\mathbf{1}^\top}{n}).$
Compute the (eigenvector, eigenvalue) pairs $(e_p, \lambda_p)$ of $K$ by Power Iteration algorithm.
    $K \leftarrow \sum_{p=1}^{n} \lambda_p e_p e_p^\top.$
Normalize the eigenvectors to get the parameters $\alpha^{(i)}$.
    $\alpha^{(i)} \leftarrow \frac{1}{\sqrt{\lambda_i}} e_i.$
Output the $i$-th kernel principal component
    $f^{(i)} \leftarrow \sum_{j=1}^{n} \alpha_j^{(i)} k(x_j, \cdot).$
```

The next theorem [2] shows that the eigenvalues of the covariance operator and the ones of the centered Gram matrix coincide. It also gives the relationship between the eigenfunctions of the covariance operator and the eigenvectors of the centered Gram matrix.

Theorem 36. *Let $(\mathcal{X}, \Omega, \mathbb{P})$ be a probability space, $\mathcal{H}$ be a separable RKHS with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, X be a $\mathcal{X}$-valued random variable and $\phi(x) := k(x, \cdot)$ be the feature map such that $\mathbb{E}\, \|\phi(X)\|_{\mathcal{H}}^2 < \infty$ and $\mathbb{E}[\phi(X)] = 0$. Let Σ_X be the covariance operator and $G : L_2(\mathbb{P}) \to L_2(\mathbb{P})$ be the integral operator defined as*

$$(Gf)(t) := \mathbb{E}[f(X)\langle \phi(X), \phi(t)\rangle_{\mathcal{H}}] = \int f(x)\langle \phi(x), \phi(t)\rangle_{\mathcal{H}} d\mathbb{P}(x), \quad \text{for all } f \in L_2(\mathbb{P}) \text{ and } t \in \mathcal{X}.$$

Then G is a Hilbert-Schmidt, positive self-adjoint operator, and

$$\lambda(G) = \lambda(\Sigma_X).$$

In particular, G is a trace-class operator, i.e., $\mathbf{Tr}(G) < \infty$, *and* $\mathbf{Tr}(G) = \sum_{i \geq 1} \lambda_i(G) = \mathbb{E}\, \|\phi(X)\|_{\mathcal{H}}^2$.

Proof We organize the proof in several steps.

Step 1. Prove G is Hilbert-Schmidt. According to Cauchy-Schwartz inequality, we have $k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}} \leq \|\phi(x)\|_{\mathcal{H}} \|\phi(y)\|_{\mathcal{H}}$ so that

$$\mathbb{E}_{X,Y}[k^2(X, Y)] \leq \left[\mathbb{E}\, \|\phi(X)\|_{\mathcal{H}}^2\right]^2 := C < \infty,$$

where the last inequality follows from the assumption that $\mathbb{E}\, \|\phi(X)\|_{\mathcal{H}}^2 < \infty$. It follows that $\phi(x) = k(x, \cdot) \in L_2(\mathbb{P})$ for any $x \in \mathcal{X}$, and thus

$$(Gf)(t) = \int f(x)k(x, t)d\mathbb{P}(x) = \langle f, k(t, \cdot)\rangle_{L_2(\mathbb{P})} = \langle f, \phi(t)\rangle_{L_2(\mathbb{P})}. \tag{17}$$

Let $\{f_i\}_{i\in I}$ be an orthonormal basis of $L_2(\mathbb{P})$, then we obtain

$$\begin{aligned}\|G\|^2_{\mathrm{HS}} &= \sum_{i\in I}\|Gf_i\|^2_{L_2(\mathbb{P})} = \sum_{i\in I}\mathbb{E}_Y[(Gf_i)(Y)^2] \overset{(17)}{=} \sum_{i\in I}\mathbb{E}_Y\langle f_i,\phi(Y)\rangle^2_{L_2(\mathbb{P})}\\ &\overset{\text{C-S}}{\le} \sum_{i\in I}\mathbb{E}_Y\Big[\,\|f_i\|^2_{L_2(\mathbb{P})}\,\|\phi(Y)\|^2_{L_2(\mathbb{P})}\,\Big] = \sum_{i\in I}\mathbb{E}_Y\Big[\,\|f_i\|^2_{L_2(\mathbb{P})}\,\mathbb{E}_X[k^2(X,Y)]\Big] = C\sum_{i\in I}\|f_i\|^2_{L_2(\mathbb{P})} < \infty,\end{aligned}$$

and thus G is Hilbert-Schmidt.

Step 2. Characterize G and Σ_X by a continuous linear operator T. Consider a linear map defined as $(Th)(x) = \langle h,\phi(x)\rangle_{\mathcal{H}}$ for any $h\in\mathcal{H}$ and $x\in\mathcal{X}$. Note that, by Cauchy-Schwartz inequality, $\|Th\|^2_{L_2(\mathbb{P})} = \mathbb{E}\{[(Th)(X)]^2\} = \mathbb{E}[\langle h,\phi(X)\rangle_{\mathcal{H}}{}^2] \le \|h\|^2_{\mathcal{H}}\,\mathbb{E}\,\|\phi(X)\|^2_{\mathcal{H}} < \infty$, so $Th\in L_2(\mathbb{P})$ for any $h\in\mathcal{H}$. That is, T is a linear operator mapping from $\mathcal{H}$ to $L_2(P)$. Moreover, T is continuous as, by the same argument, $\|Th_1 - Th_2\|^2_{L_2(\mathbb{P})} \le \|h_1 - h_2\|^2_{\mathcal{H}}\,\mathbb{E}\,\|\phi(X)\|^2_{\mathcal{H}}$. Hence, T has a continuous adjoint T^*. Next, we obtain a closed-form representation for T^*. Let $f\in L_2(\mathbb{P})$, then

$$\mathbb{E}\,\|f(X)\phi(X)\|_{\mathcal{H}} \overset{\text{C-S}}{\le} [\mathbb{E}\,\|f(X)\|^2_{\mathcal{H}}\,\mathbb{E}\,\|\phi(X)\|^2_{\mathcal{H}}]^{1/2} < \infty.$$

This implies that $\mathbb{E}[f(X)\phi(X)]\in\mathcal{H}$ is well-defined according to Riesz lemma. Furthermore, for all $h\in\mathcal{H}$, we have $\langle T^*f,h\rangle_{\mathcal{H}} = \langle f,Th\rangle_{L_2(\mathbb{P})} = \mathbb{E}[f(X)(Th)(X)] = \mathbb{E}[\langle h, f(X)\phi(X)\rangle_{\mathcal{H}}]$, and thus $T^* = \mathbb{E}[f(X)\phi(X)]$.

Now we show that $\Sigma_X = T^*T$ and $G = TT^*$. By definition, for all $h,h'\in\mathcal{H}$, $\langle h, T^*Th'\rangle_{\mathcal{H}} = \langle Th, Th'\rangle_{\mathcal{H}_{L_2(\mathbb{P})}} = \mathbb{E}[\langle h,\phi(X)\rangle_{\mathcal{H}}\langle h',\phi(X)\rangle_{\mathcal{H}}] = \mathbb{E}[h(X)h'(X)]$. Thus, by the uniqueness of the covariance operator, we get $\Sigma_X = T^*T$. Similarly,

$$(TT^*f)(x) \overset{\text{def. of } T}{=} \langle T^*f,\phi(x)\rangle_{\mathcal{H}} \overset{\text{prop. of } T^*}{=} \mathbb{E}[\langle f(X)\phi(X),\phi(x)\rangle_{\mathcal{H}}] = \int_{\mathcal{X}} f(y)\langle\phi(y),\phi(x)\rangle_{\mathcal{H}}d\mathbb{P}(y),$$

which implies $G = TT^*$ and thus G is positve self-adjoint.

Step 3. Show that nonzero eigenvalues of TT^* and T^*T coincide. Let $E_\mu(A) := \{x : Ax = \mu x\}$ be the eigenspace of the operator A associated with the eigenvalue μ. Let $\lambda > 0$ be a positive eigenvalue of $G = TT^*$ and f an associated eigenfunction. Then we have $(T^*T)T^*f = T^*(TT^*)f = \lambda T^*f$. This shows that $T^*E_\lambda(TT^*) \subset E_\lambda(T^*T)$. Applying T to both sides of this inclusion yields

$$TT^*E_\lambda(TT^*) = E_\lambda(TT^*) \subset TE_\lambda(T^*T).$$

Similarly, we have $TE_\lambda(T^*T) \subset E_\lambda(TT^*)$ and thus

$$E_\lambda(TT^*) \subset TE_\lambda(T^*T) \subset E_\lambda(TT^*).$$

This implies $E_\lambda(TT^*) = TE_\lambda(T^*T)$, and, analogously, $E_\lambda(T^*T) = T^*E_\lambda(TT^*)$. Therefore, $\dim(E_\lambda(TT^*)) = \dim(E_\lambda(T^*T))$, and it follows that λ is also an eigenvalue for Σ_X with the same multiplicity. That concludes the proof. ■