Convergence of Eigenspaces in Kernel Principal Component Analysis

Shixin Wang

Advanced machine learning

April 19, 2016

Outline

1. Motivation
2. Result
3. Summary

Outline

1. Motivation
2. Result
3. Summary

Kernel Principal Component Analysis

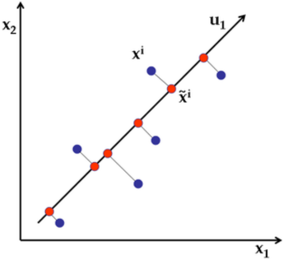
- PCA: To find the most relevant lower-dimension projection of data.
- KPCA: Extend PCA to data mapped in a kernel feature space.
- Assumption: Target dimensionality of the projected data is fixed: D.
- Objective: To find the span S_D of the first D eigenvectors of the covariance matrix.

How reliable is our estimation?

- Problem: The true covariance matrix is no known and has to be estimated from the available data.
- Question: How reliable is the D-eigenspace $\hat{S}_D$ of the empirical covariance matrix compared to the D-eigenspace S_D of the true covariance matrix?
  1. The average reconstruction error of $\hat{S}_D$ converges to the reconstruction error of S_D. (Blanchard et al 2004, Shawe-Taylor 2005)
  2. But this does not mean $\hat{S}_D$ converges to S_D! since different subspaces can have a very similar reconstruction error.

Why analyze the behavior of $\hat{S}_D$?

- PCA or KPCA is often used merely as a preprocessing step, so the behavior of $\hat{S}_D$ is more important than just reconstruction error.
- We want to use u_1 in the future, so we need to show that $\hat{x}_i$ converges to the true ones, rather than only norm of $x_i - \hat{x}_i$.



Outline

Notation

- X: interest variable taking values in measurable space $\mathcal{X}$, with distribution P.
- $\varphi(x) = k(x, \cdot)$: feature mapping of $x \in \mathcal{X}$ into a reproduction kernel Hilbert space $\mathcal{H}$
- D: target dimensionality of projected data
- C: covariance matrix of variable $\varphi(X)$
- $\lambda_1 > \lambda_2 > \ldots$: simple nonzero eigenvalues of C
- $\phi_1, \phi_2 \ldots$: associated eigenvectors
- C_n: empirical covariance matrix

Notation

- $S_D = span\{\phi_1, ...\phi_D\}$: D-dimensional subspace of $\mathcal{H}$ such that the projection of $\varphi(X)$ on S_D has maximum average squared norm
- $\hat{S}_D$: subspace spanned by the first D eigenvectors of C_n.
- P_{S_D}: the orthogonal projectors of X on S_D
- $P_{\hat{S}_D}$: the orthogonal projectors of X on $\hat{S}_D$

First Bound

Tow steps to obtain the first bound:

1. A non-asympotic bound on the (Hilbert-Schmidt) norm of the difference between the empirical and the true covariance operators
2. An operator perturbation result bounding the difference between spectral projectors of two operators by the norm of their difference.

Lemma 1

Lemma

Supposing that $\sup_{x\in\mathcal{X}} k(x,x) \leq M$*, with probability greater than* $1-e^{-\xi}$*,*

$$\|C_n - C\| \leq \frac{2M}{\sqrt{n}}(1+\sqrt{\frac{\xi}{2}})$$

Proof of Lemma 1

Proof.

Theorem (Bounded Differences Inequality)

Suppose that $X_1, \dots, X_n \in \mathcal{X}$ *are independent, and* $f : \mathcal{X}^n \to R$*,*
Let $c_1, \dots, c_n$ *satisfy*

$$\sup_{x_1,\dots,x_n,x_i^{'}} \|f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x_i^{'},x_{i+1},\dots,x_n)\| \le c_i,$$

$\forall i = 1, \dots, n$ *Then,*

$$P(f - E[f] \ge t) \le exp(\frac{-2t^2}{\sum_{i=1}^n c_i^2})$$

□

Proof of Lemma 1

Proof.

$$\|C_n - C\| = \|\frac{1}{n}\sum_{i=1}^{n} C_{X_i} - E[C_X]\|$$

$$\|C_X\| = \|\varphi(X) \otimes \varphi(X)^*\| = k(X, X) \leq M$$

Here $c_i = \frac{2M}{n}$, $t = 2M\sqrt{\frac{\xi}{2n}}$, then we get

$$P\left\{\|C_n - C\| - E[\|C_n - C\|] \geq 2M\sqrt{\frac{\xi}{2n}}\right\}$$

$$\leq exp(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}) = exp(\frac{-\frac{4M^2\xi}{n}}{4M^2/n}) = e^{-\xi}$$

Theorem 2

Theorem 2 (Simplified Version of [7], Theorem 5.2) *Let A be a symmetric positive Hilbert-Schmidt operator of the Hilbert space $\mathcal{H}$ with simple positive eigenvalues $\lambda_1 > \lambda_2 > \ldots$ For an integer r such that $\lambda_r > 0$, let $\widetilde{\delta}_r = \delta_r \wedge \delta_{r-1}$ where $\delta_r = \frac{1}{2}(\lambda_r - \lambda_{r+1})$. Let $B \in HS(\mathcal{H})$ be another symmetric operator such that $\|B\| < \widetilde{\delta}_r/2$ and $(A+B)$ is still a positive operator with simple nonzero eigenvalues.*

Let $P_r(A)$ (resp. $P_r(A+B)$) denote the orthogonal projector onto the subspace spanned by the r-th eigenvector of A (resp. $(A+B)$). Then, these projectors satisfy:

$$\|P_r(A) - P_r(A+B)\| \leq \frac{2\|B\|}{\widetilde{\delta}_r}.$$

$$\left\| P_{S_D} - P_{\widehat{S}_D} \right\| \leq \left(\sum_{r=1}^{D} \widetilde{\delta}_r^{-1} \right) \frac{4M}{\sqrt{n}} \left(1 + \sqrt{\frac{\xi}{2}} \right)$$

Improved bound

Theorem 3 *Let A be a symmetric positive Hilbert-Schmidt operator of the Hilbert space $\mathcal{H}$ with simple nonzero eigenvalues $\lambda_1 > \lambda_2 > \ldots$ Let $D > 0$ be an integer such that $\lambda_D > 0$, $\delta_D = \frac{1}{2}(\lambda_D - \lambda_{D+1})$. Let $B \in HS(\mathcal{H})$ be another symmetric operator such that $||B|| < \delta_D/2$ and $(A + B)$ is still a positive operator. Let $P^D(A)$ (resp. $P^D(A + B)$) denote the orthogonal projector onto the subspace spanned by the first D eigenvectors A (resp. $(A + B)$). Then these satisfy:*

$$||P^D(A) - P^D(A + B)|| \leq \frac{||B||}{\delta_D}. \tag{1}$$

Improved bound

Theorem 4 *Assume that* $\sup_{x\in\mathcal{X}} k(x,x) \leq M$. *Let* $S_D, \widehat{S}_D$ *be the subspaces spanned by the first* D *eigenvectors of* C, *resp.* C_n *defined earlier. Denoting* $\lambda_1 > \lambda_2 > \ldots$ *the eigenvalues of* C, *if* $D > 0$ *is such that* $\lambda_D > 0$, *put* $\delta_D = \frac{1}{2}(\lambda_D - \lambda_{D+1})$ *and*

$$B_D = \frac{2M}{\delta_D}\left(1 + \sqrt{\frac{\xi}{2}}\right).$$

Then provided that $n \geq B_D^2$, *the following bound holds with probability at least* $1 - e^{-\xi}$:

$$\left\| P_{S_D} - P_{\widehat{S}_D} \right\| \leq \frac{B_D}{\sqrt{n}}. \qquad (2)$$

Outline

Summary

- Provide finite sample size confidence bounds of the eigenspaces of Kernel-PCA
- Prove a bound in which the complexity factor for estimating the eigenspace S_D by its empirical counterpart depends only on the inverse gap between the D-th and $D+1$-th eigenvalues
- Restriction: Assume that the eigenvalues to be simple.