Approximate Kernel PCA with Random Features

(Computational vs. Statistical Tradeoff)

Bharath K. Sriperumbudur

Department of Statistics, Pennsylvania State University

Journées de Statistique
Paris
May 28, 2018

(Joint work with Nicholas Sterge, PSU)

Supported by National Science Foundation (DMS-1713011)

Outline

- Principal Component Analysis (PCA)
- Kernel PCA
- Approximation methods
- Kernel PCA with Random Features
- Computational vs. statistical trade off

Principal Component Analysis (PCA) (Pearson, 1901)

- Suppose $X \sim \mathbb{P}$ be a random variable in $\mathbb{R}^d$ with mean μ and covariance matrix Σ.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that

$$\mathrm{Var}[\langle w, X\rangle_2] = \langle w, \Sigma w\rangle_2$$

  is maximized.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that

$$\mathbb{E}\|(X-\mu) - \langle w, (X-\mu)\rangle_2 w\|_2^2$$

  is minimized.
- The formulations are equivalent and the solution is the eigenvector of the covariance matrix Σ corresponding to the largest eigenvalue.
- Can be generalized to multiple directions (find a subspace...).
- Applications: dimensionality reduction.

Principal Component Analysis (PCA) (Pearson, 1901)

- Suppose $X \sim \mathbb{P}$ be a random variable in $\mathbb{R}^d$ with mean μ and covariance matrix Σ.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that
$$\mathrm{Var}[\langle w, X\rangle_2] = \langle w, \Sigma w\rangle_2$$
is maximized.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that
$$\mathbb{E}\|(X-\mu) - \langle w, (X-\mu)\rangle_2 w\|_2^2$$
is minimized.
- The formulations are equivalent and the solution is the eigenvector of the covariance matrix Σ corresponding to the largest eigenvalue.
- Can be generalized to multiple directions (find a subspace...).
- Applications: dimensionality reduction.

Principal Component Analysis (PCA) (Pearson, 1901)

- Suppose $X \sim \mathbb{P}$ be a random variable in $\mathbb{R}^d$ with mean μ and covariance matrix Σ.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that
$$\mathrm{Var}[\langle w, X\rangle_2] = \langle w, \Sigma w\rangle_2$$
is maximized.
- Find a direction $w \in \{v : \|v\|_2 = 1\}$ such that
$$\mathbb{E}\|(X-\mu) - \langle w, (X-\mu)\rangle_2 w\|_2^2$$
is minimized.
- The formulations are equivalent and the solution is the eigenvector of the covariance matrix Σ corresponding to the largest eigenvalue.
- Can be generalized to multiple directions (find a subspace...).
- Applications: dimensionality reduction.

Kernel PCA

- Nonlinear generalization of PCA (Schölkopf et al., 1998).
- $X \mapsto \Phi(X)$ through the feature map Φ and apply PCA.
- The choice of Φ determines the degree of information we capture about X.
- Suppose $\Phi(X) = (1, X, X^2, X^3, \ldots)$, then the covariance of $\Phi(X)$ captures the higher order moments of X.
- Φ is not explicitly specified but implicitly specified through a positive definite kernel function, $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$.

Kernel PCA: Functional Version

- Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\text{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \text{Var}[f(X)]$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (evaluational functionals $f \mapsto f(x)$ are bounded for all $x \in \mathcal{X}$) of real-valued functions (Aronszajn, 1950).

- $\exists$ unique $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{H}$ and $f(x) = \langle k(\cdot, x), f\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, $x \in \mathcal{X}$.

- k is called the reproducing kernel of $\mathcal{H}$ as

$$k(x, y) = \langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, y)}_{\Phi(y)} \rangle_{\mathcal{H}}$$

and is symmetric and positive definite. In fact, the converse is also true (Moore-Aronszajn Theorem).

Kernel PCA: Functional Version

- Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\mathrm{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \mathrm{Var}[f(X)]$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (evaluational functionals $f \mapsto f(x)$ are bounded for all $x \in \mathcal{X}$) of real-valued functions (Aronszajn, 1950).

- $\exists$ unique $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{H}$ and $f(x) = \langle k(\cdot, x), f\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, $x \in \mathcal{X}$.

- k is called the reproducing kernel of $\mathcal{H}$ as

$$k(x, y) = \langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, y)}_{\Phi(y)} \rangle_{\mathcal{H}}$$

  and is symmetric and positive definite. In fact, the converse is also true (Moore-Aronszajn Theorem).

Kernel PCA: Functional Version

- Find $f \in \{g \in \mathcal{H} : \|g\|_{\mathcal{H}} = 1\}$ such that $\mathrm{Var}[f(X)]$ is maximized, i.e.,

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \mathrm{Var}[f(X)]$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (evaluational functionals $f \mapsto f(x)$ are bounded for all $x \in \mathcal{X}$) of real-valued functions (Aronszajn, 1950).

- $\exists$ unique $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(\cdot, x) \in \mathcal{H}$ for all $x \in \mathcal{H}$ and $f(x) = \langle k(\cdot, x), f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, $x \in \mathcal{X}$.

- k is called the reproducing kernel of $\mathcal{H}$ as

$$k(x, y) = \langle \underbrace{k(\cdot, x)}_{\Phi(x)}, \underbrace{k(\cdot, y)}_{\Phi(y)} \rangle_{\mathcal{H}}$$

  and is symmetric and positive definite. In fact, the converse is also true (Moore-Aronszajn Theorem).

RKHS

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$ (linear span of kernel functions)
- k controls the properties of $f \in \mathcal{H}$.
- If k satisfies $(*)$, then every $f \in \mathcal{H}$ satisfies $(*)$, where $(*)$ is
  - boundedness
  - continuity
  - measurability
  - integrability
  - differentiability

RKHS

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$ (linear span of kernel functions)
- k controls the properties of $f \in \mathcal{H}$.
- If k satisfies $(*)$, then every $f \in \mathcal{H}$ satisfies $(*)$, where $(*)$ is
  - boundedness
  - continuity
  - measurability
  - integrability
  - differentiability

RKHS

Kernels $\Leftrightarrow$ Positive definite & symmetric functions $\Leftrightarrow$ RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$ (linear span of kernel functions)
- k controls the properties of $f \in \mathcal{H}$.
- If k satisfies $(*)$, then every $f \in \mathcal{H}$ satisfies $(*)$, where $(*)$ is
  - boundedness
  - continuity
  - measurability
  - integrability
  - differentiability

Kernel PCA

- Kernel PCA generalizes linear PCA.
- $k(x,y) = \langle x,y \rangle_2$, $x,y \in \mathbb{R}^d$: $\mathcal{H}$ is isometrically isomorphic to $\mathbb{R}^d$.

$$f(x) = \langle w_f, x \rangle_2, \ \forall f \in \mathcal{H}$$

Kernel PCA

- Kernel PCA generalizes linear PCA.
- $k(x, y) = \langle x, y \rangle_2$, $x, y \in \mathbb{R}^d$: $\mathcal{H}$ is isometrically isomorphic to $\mathbb{R}^d$.

$$f(x) = \langle w_f, x \rangle_2, \ \forall f \in \mathcal{H}$$

Kernel PCA

- Using the reproducing property $f(X) = \langle f, \underbrace{k(\cdot, X)}_{\Phi(X)} \rangle_{\mathcal{H}}$, we obtain

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \Sigma f \rangle_{\mathcal{H}}.$$

- $$\Sigma := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x) - \mu_{\mathbb{P}} \otimes_{\mathcal{H}} \mu_{\mathbb{P}}$$

  is the covariance operator (self-adjoint, positive and trace class) on $\mathcal{H}$ and

  $$\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$$

  is the mean element.

- Spectral theorem:

  $$\Sigma = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$$

  where I is either countable ($\lambda_i \to 0$ as $i \to \infty$) or finite.

- Similar to PCA, the solution is the eigenfunction corresponding to the largest eigenvalue of Σ.

Kernel PCA

- Using the reproducing property $f(X) = \langle f, \underbrace{k(\cdot, X)}_{\Phi(X)} \rangle_{\mathcal{H}}$, we obtain

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \Sigma f \rangle_{\mathcal{H}}.$$

- $$\Sigma := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x) - \mu_{\mathbb{P}} \otimes_{\mathcal{H}} \mu_{\mathbb{P}}$$

  is the covariance operator (self-adjoint, positive and trace class) on $\mathcal{H}$ and

  $$\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$$

  is the mean element.

- Spectral theorem:

  $$\Sigma = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$$

  where I is either countable ($\lambda_i \to 0$ as $i \to \infty$) or finite.

- Similar to PCA, the solution is the eigenfunction corresponding to the largest eigenvalue of Σ.

Kernel PCA

- Using the reproducing property $f(X) = \langle f, \underbrace{k(\cdot, X)}_{\Phi(X)} \rangle_{\mathcal{H}}$, we obtain

$$f^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \Sigma f \rangle_{\mathcal{H}}.$$

- $$\Sigma := \int_{\mathcal{X}} k(\cdot, x) \otimes_{\mathcal{H}} k(\cdot, x)\, d\mathbb{P}(x) - \mu_{\mathbb{P}} \otimes_{\mathcal{H}} \mu_{\mathbb{P}}$$

  is the covariance operator (self-adjoint, positive and trace class) on $\mathcal{H}$ and

$$\mu_{\mathbb{P}} := \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$$

  is the mean element.

- Spectral theorem:

$$\Sigma = \sum_{i \in I} \lambda_i \phi_i \otimes_{\mathcal{H}} \phi_i$$

  where I is either countable ($\lambda_i \to 0$ as $i \to \infty$) or finite.

- Similar to PCA, the solution is the eigenfunction corresponding to the largest eigenvalue of Σ.

Empirical Kernel PCA

In practice, $\mathbb{P}$ is unknown but have access to $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

- $$\hat{f}^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f \rangle_{\mathcal{H}},$$

  where

  $$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) - \mu_n \otimes_{\mathcal{H}} \mu_n$$

  is the empirical covariance operator (symmetric, positive and trace class) on $\mathcal{H}$ and

  $$\mu_n := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i).$$

- Spectral theorem:

  $$\hat{\Sigma} = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i.$$

- $\hat{f}^*$ is the eigenfunction corresponding to the largest eigenvalue of $\hat{\Sigma}$.

Empirical Kernel PCA

In practice, $\mathbb{P}$ is unknown but have access to $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

- $$\hat{f}^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f \rangle_{\mathcal{H}},$$

  where

  $$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) - \mu_n \otimes_{\mathcal{H}} \mu_n$$

  is the empirical covariance operator (symmetric, positive and trace class) on $\mathcal{H}$ and

  $$\mu_n := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i).$$

- Spectral theorem:

  $$\hat{\Sigma} = \sum_{i=1}^{n} \hat{\lambda}_i \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i.$$

- $\hat{f}^*$ is the eigenfunction corresponding to the largest eigenvalue of $\hat{\Sigma}$.

Empirical Kernel PCA

In practice, $\mathbb{P}$ is unknown but have access to $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

- $$\hat{f}^* = \arg \sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f \rangle_{\mathcal{H}},$$

  where

  $$\hat{\Sigma} := \frac{1}{n} \sum_{i=1}^n k(\cdot, X_i) \otimes_{\mathcal{H}} k(\cdot, X_i) - \mu_n \otimes_{\mathcal{H}} \mu_n$$

  is the empirical covariance operator (symmetric, positive and trace class) on $\mathcal{H}$ and

  $$\mu_n := \frac{1}{n} \sum_{i=1}^n k(\cdot, X_i).$$

- Spectral theorem:

  $$\hat{\Sigma} = \sum_{i=1}^n \hat{\lambda}_i \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i.$$

- $\hat{f}^*$ is the eigenfunction corresponding to the largest eigenvalue of $\hat{\Sigma}$.

Empirical Kernel PCA: Representer Theorem

- Since $\hat{\Sigma}$ is an infinite dimensional operator, we have to solve an infinite dimensional eigen system,

$$\hat{\Sigma}\hat{\phi}_i = \hat{\lambda}_i\hat{\phi}_i.$$

- Consider

$$\sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f\rangle_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^{n}\langle f, k(\cdot, X_i)\rangle^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)\right\rangle_{\mathcal{H}}^2.$$

- Clearly $f^* \in \text{span}\{k(\cdot, X_i) : i = 1, \ldots, n\}$, i.e.,

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i).$$

- $\sup\left\{\langle f, \hat{\Sigma} f\rangle_{\mathcal{H}} : \|f\|_{\mathcal{H}} = 1\right\} = \sup\left\{\boldsymbol{\alpha}^\top \mathbf{K}\mathbf{H}_n\mathbf{K}\boldsymbol{\alpha} : \boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha} = 1\right\}$, i.e.,

$$\mathbf{K}\mathbf{H}_n\mathbf{K}\boldsymbol{\alpha} = \lambda\mathbf{K}\boldsymbol{\alpha}.$$

- Requires $\mathbf{K}$ to be invertible.

Empirical Kernel PCA: Representer Theorem

- Since $\hat{\Sigma}$ is an infinite dimensional operator, we have to solve an infinite dimensional eigen system,

$$\hat{\Sigma}\hat{\phi}_i = \hat{\lambda}_i\hat{\phi}_i.$$

- Consider

$$\sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f\rangle_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^{n}\langle f, k(\cdot, X_i)\rangle^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)\right\rangle_{\mathcal{H}}^2.$$

- Clearly $f^* \in \text{span}\{k(\cdot, X_i) : i = 1, \dots, n\}$, i.e.,

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i).$$

- $\sup\left\{\langle f, \hat{\Sigma} f\rangle_{\mathcal{H}} : \|f\|_{\mathcal{H}} = 1\right\} = \sup\left\{\boldsymbol{\alpha}^\top \mathbf{K}\mathbf{H}_n\mathbf{K}\boldsymbol{\alpha} : \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} = 1\right\}$, i.e.,

$$\mathbf{K}\mathbf{H}_n\mathbf{K}\boldsymbol{\alpha} = \lambda\mathbf{K}\boldsymbol{\alpha}.$$

- Requires $\mathbf{K}$ to be invertible.

Empirical Kernel PCA: Representer Theorem

- Since $\hat{\Sigma}$ is an infinite dimensional operator, we have to solve an infinite dimensional eigen system,

$$\hat{\Sigma}\hat{\phi}_i = \hat{\lambda}_i\hat{\phi}_i.$$

- Consider

$$\sup_{\|f\|_{\mathcal{H}}=1} \langle f, \hat{\Sigma} f\rangle_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \frac{1}{n}\sum_{i=1}^{n} \langle f, k(\cdot, X_i)\rangle^2 - \left\langle f, \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_i)\right\rangle_{\mathcal{H}}^2.$$

- Clearly $f^* \in \mathrm{span}\{k(\cdot, X_i) : i = 1, \ldots, n\}$, i.e.,

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, X_i).$$

- $\sup\left\{\langle f, \hat{\Sigma} f\rangle_{\mathcal{H}} : \|f\|_{\mathcal{H}} = 1\right\} = \sup\left\{\boldsymbol{\alpha}^\top \mathbf{K}\mathbf{H}_n\mathbf{K}\boldsymbol{\alpha} : \boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} = 1\right\}$, i.e.,

$$\mathbf{K}\mathbf{H}_n\mathbf{K}\boldsymbol{\alpha} = \lambda\mathbf{K}\boldsymbol{\alpha}.$$

- Requires $\mathbf{K}$ to be invertible.

Empirical Kernel PCA

- In classical PCA, $X_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ is represented as

$$(\langle X_i - \overline{\mu}, \hat{w}_1 \rangle_2, \ldots, \langle X_i - \overline{\mu}, \hat{w}_\ell \rangle_2) \in \mathbb{R}^\ell$$

with $\ell \leq d$ where $(\hat{w}_i)_{i=1}^\ell$ are the eigenvectors of $\hat{\Sigma}$ corresponding to the top-ℓ eigenvalues.

- In kernel PCA, $X_i \in \mathcal{X}$, $i = 1, \ldots, n$ is represented as

$$\left(\left\langle k(\cdot, X_i) - \mu_n, \hat{\phi}_1 \right\rangle_{\mathcal{H}}, \ldots, \left\langle k(\cdot, X_i) - \mu_n, \hat{\phi}_\ell \right\rangle_{\mathcal{H}} \right) \in \mathbb{R}^\ell$$

with $\ell \leq n$ where $(\hat{\phi}_i)_{i=1}^\ell$ are the eigenfunctions of $\hat{\Sigma}$ corresponding to the top-ℓ eigenvalues.

Empirical Kernel PCA

- In classical PCA, $X_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ is represented as

$$(\langle X_i - \overline{\mu}, \hat{w}_1 \rangle_2, \ldots, \langle X_i - \overline{\mu}, \hat{w}_\ell \rangle_2) \in \mathbb{R}^\ell$$

  with $\ell \leq d$ where $(\hat{w}_i)_{i=1}^{\ell}$ are the eigenvectors of $\hat{\Sigma}$ corresponding to the top-ℓ eigenvalues.

- In kernel PCA, $X_i \in \mathcal{X}$, $i = 1, \ldots, n$ is represented as

$$\left(\left\langle k(\cdot, X_i) - \mu_n, \hat{\phi}_1 \right\rangle_{\mathcal{H}}, \ldots, \left\langle k(\cdot, X_i) - \mu_n, \hat{\phi}_\ell \right\rangle_{\mathcal{H}} \right) \in \mathbb{R}^\ell$$

  with $\ell \leq n$ where $(\hat{\phi}_i)_{i=1}^{\ell}$ are the eigenfunctions of $\hat{\Sigma}$ corresponding to the top-ℓ eigenvalues.

Summary

- The direct formulation requires the knowledge of feature map Φ (and of course $\mathcal{H}$) and these could be infinite dimensional.

$$\hat{\Sigma}\hat{\phi}_i = \hat{\lambda}_i\hat{\phi}_i.$$

- The alternate formulation is entirely determined by kernel evaluations, Gram matrix. But poor scalability: $O(n^3)$.

$$\hat{\phi}_i = \sum_{j=1}^{n} \alpha_{i,j} k(\cdot, X_j),$$

where $\boldsymbol{\alpha}_i$ satisfies

$$\mathbf{H}_n \mathbf{K} \boldsymbol{\alpha}_i = \lambda_i \boldsymbol{\alpha}_i.$$

Approximation Schemes

- Incomplete Cholesky factorization (e.g., Fine and Scheinberg, 2001)
- Sketching (Yang et al., 2015)
- Sparse greedy approximation (Smola and Schölkopf, 2000)
- Nyström method (e.g., Williams and Seeger, 2001)
- Random Fourier features (e.g., Rahimi and Recht, 2008a), ...

Random Fourier Approximation

- $\mathcal{X} = \mathbb{R}^d$; k be continuous and translation-invariant, i.e., $k(x,y) = \psi(x-y)$.
- Bochner's theorem: ψ is positive definite if and only if

$$k(x,y) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle \omega, x-y\rangle_2}\, d\Lambda(\omega),$$

where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.

  - k is symmetric and therefore Λ is a "symmetric" measure on $\mathbb{R}^d$.

- Therefore

$$k(x,y) = \int_{\mathbb{R}^d} \cos(\langle \omega, x-y\rangle_2)\, d\Lambda(\omega).$$

Random Fourier Approximation

- $\mathcal{X} = \mathbb{R}^d$; k be continuous and translation-invariant, i.e., $k(x, y) = \psi(x - y)$.
- Bochner's theorem: ψ is positive definite if and only if

$$k(x, y) = \int_{\mathbb{R}^d} e^{\sqrt{-1}\langle \omega, x-y \rangle_2} \, d\Lambda(\omega),$$

  where Λ is a finite non-negative Borel measure on $\mathbb{R}^d$.
  - k is symmetric and therefore Λ is a "symmetric" measure on $\mathbb{R}^d$.
- Therefore

$$k(x, y) = \int_{\mathbb{R}^d} \cos(\langle \omega, x - y \rangle_2) \, d\Lambda(\omega).$$

Random Feature Approximation

(Rahimi and Recht, 2008a): Draw $(\omega_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$.

$$k_m(x,y) = \frac{1}{m}\sum_{j=1}^m \cos\left(\langle \omega_j, x-y\rangle_2\right) = \langle \Phi_m(x), \Phi_m(y)\rangle_{\mathbb{R}^{2m}},$$

$$\approx k(x,y) = \underbrace{\langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}}_{\langle \Phi(x),\Phi(y)\rangle_{\mathcal{H}}}$$

where

$$\Phi_m(x) = \frac{1}{\sqrt{m}}\left(\overbrace{\cos(\langle\omega_1,x\rangle_2)}^{\varphi_1(x)},\ldots,\cos(\langle\omega_m,x\rangle_2),\sin(\langle\omega_1,x\rangle_2),\ldots,\sin(\langle\omega_m,x\rangle_2)\right).$$

Idea: Apply PCA to $\Phi_m(x)$.

Random Feature Approximation

(Rahimi and Recht, 2008a): Draw $(\omega_j)_{j=1}^m \overset{i.i.d.}{\sim} \Lambda$.

$$
\begin{aligned}
k_m(x,y) = \frac{1}{m}\sum_{j=1}^m \cos\left(\langle \omega_j, x - y\rangle_2\right) &= \langle \Phi_m(x), \Phi_m(y)\rangle_{\mathbb{R}^{2m}}, \\
&\approx k(x,y) = \underbrace{\langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{H}}}_{\langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}}
\end{aligned}
$$

where

$$
\Phi_m(x) = \frac{1}{\sqrt{m}}\left(\overbrace{\cos(\langle \omega_1, x\rangle_2)}^{\varphi_1(x)}, \ldots, \cos(\langle \omega_m, x\rangle_2), \sin(\langle \omega_1, x\rangle_2), \ldots, \sin(\langle \omega_m, x\rangle_2)\right).
$$

Idea: Apply PCA to $\Phi_m(x)$.

Approximate Kernel PCA (RF-KPCA)

- Perform linear PCA on $\Phi_m(X)$ where $X \sim \mathbb{P}$.
- Approximate empirical KPCA finds $\beta \in \mathbb{R}^m$ that solves

$$\sup_{\|\beta\|_2=1} \mathrm{Var}[\langle \beta, \Phi_m(X)\rangle_2] = \sup_{\|\beta\|_2=1} \langle \beta, \Sigma_m \beta\rangle_2$$

where $\Sigma_m := \mathbb{E}[\Phi_m(X) \otimes_2 \Phi_m(X)] - \mathbb{E}[\Phi_m(X)] \otimes_2 \mathbb{E}[\Phi_m(X)]$.

- Same as doing kernel PCA in $\mathcal{H}_m$ where

$$\mathcal{H}_m = \left\{ f = \sum_{i=1}^{m} \beta_i \varphi_i \,:\, \beta \in \mathbb{R}^m \right\}$$

is an RKHS induced by the reproducing kernel k_m w.r.t.

$$\langle f, g\rangle_{\mathcal{H}_m} = \langle \beta_f, \beta_g\rangle_2 .$$

Approximate Kernel PCA (RF-KPCA)

- Perform linear PCA on $\Phi_m(X)$ where $X \sim \mathbb{P}$.
- Approximate empirical KPCA finds $\beta \in \mathbb{R}^m$ that solves

$$\sup_{\|\beta\|_2=1} \mathrm{Var}[\langle \beta, \Phi_m(X)\rangle_2] = \sup_{\|\beta\|_2=1} \langle \beta, \Sigma_m \beta\rangle_2$$

where $\Sigma_m := \mathbb{E}[\Phi_m(X) \otimes_2 \Phi_m(X)] - \mathbb{E}[\Phi_m(X)] \otimes_2 \mathbb{E}[\Phi_m(X)]$.

- Same as doing kernel PCA in $\mathcal{H}_m$ where

$$\mathcal{H}_m = \left\{ f = \sum_{i=1}^{m} \beta_i \varphi_i \, : \, \beta \in \mathbb{R}^m \right\}$$

is an RKHS induced by the reproducing kernel k_m w.r.t.

$$\langle f, g\rangle_{\mathcal{H}_m} = \langle \beta_f, \beta_g\rangle_2 \, .$$

Empirical RF-KPCA

The empirical counterpart is obtained as:

$$\hat{\beta}_m^* = \arg \sup_{\|\beta\|_2=1} \left\langle \beta, \hat{\Sigma}_m \beta \right\rangle_2$$

where

$$\hat{\Sigma}_m := \frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i) \otimes_2 \Phi_m(X_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\right) \otimes_2 \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\right).$$

- Eigen decomposition: $\hat{\Sigma}_m = \sum_{i=1}^{m} \hat{\lambda}_{i,m} \hat{\phi}_{i,m} \otimes_2 \hat{\phi}_{i,m}$
- $\hat{\beta}_m^*$ is obtained by solving an $m \times m$ eigensystem: Complexity is $O(m^3)$.

What happens statistically?

Empirical RF-KPCA

The empirical counterpart is obtained as:

$$\hat{\beta}^*_m = \arg \sup_{\|\beta\|_2=1} \left\langle \beta, \hat{\Sigma}_m \beta \right\rangle_2$$

where

$$\hat{\Sigma}_m := \frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i) \otimes_2 \Phi_m(X_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\right) \otimes_2 \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\right).$$

- Eigen decomposition: $\hat{\Sigma}_m = \sum_{i=1}^{m} \hat{\lambda}_{i,m} \hat{\phi}_{i,m} \otimes_2 \hat{\phi}_{i,m}$
- $\hat{\beta}^*_m$ is obtained by solving an $m \times m$ eigensystem: Complexity is $O(m^3)$.

What happens statistically?

Empirical RF-KPCA

The empirical counterpart is obtained as:

$$\hat{\beta}^*_m = \arg \sup_{\|\beta\|_2=1} \left\langle \beta, \hat{\Sigma}_m \beta \right\rangle_2$$

where

$$\hat{\Sigma}_m := \frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i) \otimes_2 \Phi_m(X_i) - \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\right) \otimes_2 \left(\frac{1}{n}\sum_{i=1}^{n} \Phi_m(X_i)\right).$$

- Eigen decomposition: $\hat{\Sigma}_m = \sum_{i=1}^{m} \hat{\lambda}_{i,m} \hat{\phi}_{i,m} \otimes_2 \hat{\phi}_{i,m}$
- $\hat{\beta}^*_m$ is obtained by solving an $m \times m$ eigensystem: Complexity is $O(m^3)$.

What happens statistically?

How good is the approximation?

(S and Szabó, 2016):

$$\sup_{x,y \in \mathscr{S}} |k_m(x,y) - k(x,y)| = O_{a.s.}\left(\sqrt{\frac{\log|\mathscr{S}|}{m}}\right)$$

Optimal convergence rate

- Other results are known but they are non-optimal (Rahimi and Recht, 2008a; Sutherland and Schneider, 2015).

What happens statistically?

Kernel ridge regression: $(X_i, Y_i)_{i=1}^n \overset{iid}{\sim} \rho_{XY}$.

- $\mathcal{R}^*_{\mathbf{P}} = \inf_{f \in L^2(\rho_X)} \mathbb{E}|f(X) - Y|^2 = \mathbb{E}|f^*(X) - Y|^2$.
- Penalized risk minimization: $O(n^3)$

$$f_n = \arg \inf_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|_2^2 + \lambda \|f\|_{\mathcal{H}}^2$$

- Penalized risk minimization (approximate): $O(m^2 n)$

$$f_{m,n} = \arg \inf_{f \in \mathcal{H}_m} \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X_i)|_2^2 + \lambda \|f\|_{\mathcal{H}_m}^2$$

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.
- Similar results are derived for Nyström approximation (Bach 2013, Alaoui and Mahoney, 2015, Rudi et al., 2015).

Computational gain with no statistical loss!!

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.
- Similar results are derived for Nyström approximation (Bach 2013, Alaoui and Mahoney, 2015, Rudi et al., 2015).

Computational gain with no statistical loss!!

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.
- Similar results are derived for Nyström approximation (Bach 2013, Alaoui and Mahoney, 2015, Rudi et al., 2015).

Computational gain with no statistical loss!!

What happens statistically?

$$\underbrace{\mathcal{R}_{\mathbf{P}}(f_{m,n})}_{\mathbb{E}|f_{m,n}(X)-Y|^2} \; - \mathcal{R}^*$$

$$= \underbrace{(\mathcal{R}_{\mathbf{P}}(f_{m,n}) - \mathcal{R}_{\mathbf{P}}(f_n))}_{\text{error due to approximation}} + (\mathcal{R}_{\mathbf{P}}(f_n) - \mathcal{R}^*_{\mathbf{P}})$$

- (Rahimi and Recht, 2008b): $(m \wedge n)^{-\frac{1}{2}}$
- (Rudi and Rosasco, 2016): If $m \geq n^{\alpha}$ where $\frac{1}{2} \leq \alpha < 1$ with α depending on the properties of f^*, then $f_{m,n}$ achieves the minimax optimal rate as obtained in the case with no approximation.
- Similar results are derived for Nyström approximation (Bach 2013, Alaoui and Mahoney, 2015, Rudi et al., 2015).

Computational gain with no statistical loss!!

What happens statistically?

Two notions for PCA:

- Reconstruction error
- Convergence of eigenspaces

Reconstruction Error

- Linear PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|(X-\mu)-\sum_{i=1}^{\ell}\langle(X-\mu),\phi_i\rangle_2\phi_i\right\|_2^2$$

- Kernel PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|\tilde{k}(\cdot,X)-\sum_{i=1}^{\ell}\langle\tilde{k}(\cdot,X),\phi_i\rangle_{\mathcal{H}}\phi_i\right\|_{\mathcal{H}}^2$$

where $\tilde{k}(\cdot,x)=k(\cdot,x)-\int k(\cdot,x)\,d\mathbb{P}(x)$.

- However, the eigenfunctions of approximate empirical KPCA lie in $\mathcal{H}_m$, which is finite dimensional and not contained in $\mathcal{H}$.

Reconstruction Error

- Linear PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|(X-\mu)-\sum_{i=1}^{\ell}\langle (X-\mu),\phi_i\rangle_2\phi_i\right\|_2^2$$

- Kernel PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|\tilde{k}(\cdot,X)-\sum_{i=1}^{\ell}\langle \tilde{k}(\cdot,X),\phi_i\rangle_{\mathcal{H}}\phi_i\right\|_{\mathcal{H}}^2$$

where $\tilde{k}(\cdot,x)=k(\cdot,x)-\int k(\cdot,x)\,d\mathbb{P}(x)$.

- However, the eigenfunctions of approximate empirical KPCA lie in $\mathcal{H}_m$, which is finite dimensional and not contained in $\mathcal{H}$.

Reconstruction Error

- Linear PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|(X-\mu)-\sum_{i=1}^{\ell}\langle(X-\mu),\phi_i\rangle_2\phi_i\right\|_2^2$$

- Kernel PCA

$$\mathbb{E}_{X\sim\mathbb{P}}\left\|\tilde{k}(\cdot,X)-\sum_{i=1}^{\ell}\langle\tilde{k}(\cdot,X),\phi_i\rangle_{\mathcal{H}}\phi_i\right\|_{\mathcal{H}}^2$$

where $\tilde{k}(\cdot,x)=k(\cdot,x)-\int k(\cdot,x)\,d\mathbb{P}(x)$.

- However, the eigenfunctions of approximate empirical KPCA lie in $\mathcal{H}_m$, which is finite dimensional and not contained in $\mathcal{H}$.

Embedding to $L^2(\mathbb{P})$

What we have?

- Population eigenfunctions $(\phi_i)_{i \in I}$ of Σ: these form a subspace in $\mathcal{H}$.
- Empirical eigenfunctions $(\hat{\phi}_i)_{i=1}^n$ of $\hat{\Sigma}$: these form a subspace in $\mathcal{H}$.
- Eigenvectors after approximation, $(\hat{\phi}_{i,m})_{i=1}^m$ of $\hat{\Sigma}_m$: these form a subspace in $\mathbb{R}^m$
- We embed them in a common space before comparing. The common space is $L^2(\mathbb{P})$.
- (Inclusion operator) $\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- (Approximation operator) $\mathfrak{U} : \mathbb{R}^m \to L^2(\mathbb{P})$,

$$\boldsymbol{\alpha} \mapsto \sum_{i=1}^{m} \alpha_i \left(\varphi_i - \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x) \right)$$

Embedding to $L^2(\mathbb{P})$

What we have?

- Population eigenfunctions $(\phi_i)_{i\in I}$ of Σ: these form a subspace in $\mathcal{H}$.
- Empirical eigenfunctions $(\hat{\phi}_i)_{i=1}^n$ of $\hat{\Sigma}$: these form a subspace in $\mathcal{H}$.
- Eigenvectors after approximation, $(\hat{\phi}_{i,m})_{i=1}^m$ of $\hat{\Sigma}_m$: these form a subspace in $\mathbb{R}^m$
- We embed them in a common space before comparing. The common space is $L^2(\mathbb{P})$.

- (Inclusion operator) $\mathfrak{I} : \mathcal{H} \to L^2(\mathbb{P})$, $f \mapsto f - \int_{\mathcal{X}} f(x)\, d\mathbb{P}(x)$
- (Approximation operator) $\mathfrak{U} : \mathbb{R}^m \to L^2(\mathbb{P})$,

$$\boldsymbol{\alpha} \mapsto \sum_{i=1}^{m} \alpha_i \left(\varphi_i - \int_{\mathcal{X}} \varphi_i(x)\, d\mathbb{P}(x) \right)$$

Properties

- $\Sigma = \mathfrak{I}^*\mathfrak{I}$
- $\mathfrak{I}$ and $\mathfrak{I}^*$ are HS and Σ is trace-class
- $\Sigma_m = \mathfrak{U}^*\mathfrak{U}$
- $\mathfrak{U}$ and $\mathfrak{U}^*$ are HS and Σ_m is trace-class
- $\|\Sigma - \hat{\Sigma}\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\Sigma_m - \hat{\Sigma}_m\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{U}\mathfrak{U}^*\|_{HS} = O_{\Lambda}\left(m^{-\frac{1}{2}}\right)$

Properties

- $\Sigma = \mathfrak{I}^*\mathfrak{I}$
- $\mathfrak{I}$ and $\mathfrak{I}^*$ are HS and Σ is trace-class
- $\Sigma_m = \mathfrak{U}^*\mathfrak{U}$
- $\mathfrak{U}$ and $\mathfrak{U}^*$ are HS and Σ_m is trace-class
- $\|\Sigma - \hat{\Sigma}\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\Sigma_m - \hat{\Sigma}_m\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{U}\mathfrak{U}^*\|_{HS} = O_{\Lambda}\left(m^{-\frac{1}{2}}\right)$

Properties

- $\Sigma = \mathfrak{I}^*\mathfrak{I}$
- $\mathfrak{I}$ and $\mathfrak{I}^*$ are HS and Σ is trace-class
- $\Sigma_m = \mathfrak{U}^*\mathfrak{U}$
- $\mathfrak{U}$ and $\mathfrak{U}^*$ are HS and Σ_m is trace-class
- $\|\Sigma - \hat{\Sigma}\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\Sigma_m - \hat{\Sigma}_m\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{U}\mathfrak{U}^*\|_{HS} = O_{\Lambda}\left(m^{-\frac{1}{2}}\right)$

Properties

- $\Sigma = \mathfrak{I}^*\mathfrak{I}$
- $\mathfrak{I}$ and $\mathfrak{I}^*$ are HS and Σ is trace-class
- $\Sigma_m = \mathfrak{U}^*\mathfrak{U}$
- $\mathfrak{U}$ and $\mathfrak{U}^*$ are HS and Σ_m is trace-class
- $\|\Sigma - \hat{\Sigma}\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\Sigma_m - \hat{\Sigma}_m\|_{HS} = O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right)$
- $\|\mathfrak{I}\mathfrak{I}^* - \mathfrak{U}\mathfrak{U}^*\|_{HS} = O_{\Lambda}\left(m^{-\frac{1}{2}}\right)$

Reconstruction Error

$\left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$. Define $\tilde{k}(\cdot, x) = k(\cdot, x) - \mu_{\mathbb{P}}$ and $\tau > 0$.

- Population KPCA:

$$R_\ell = \mathbb{E}\left\|\mathfrak{I}\tilde{k}(\cdot, X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}, \mathfrak{I}\tilde{k}(\cdot, X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\Sigma - \Sigma^{1/2}\Sigma_\ell^{-1}\Sigma^{3/2}\right\|_{HS}^2 = \|\Sigma - \Sigma_\ell\|_{HS}^2 .$$

- Empirical KPCA:

$$R_{n,\ell} = \mathbb{E}\left\|\mathfrak{I}\tilde{k}(\cdot, X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}, \mathfrak{I}\tilde{k}(\cdot, X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\Sigma - \Sigma^{1/2}\hat{\Sigma}_\ell^{-1}\Sigma^{3/2}\right\|_{HS}^2$$

- Approximate Empirical KPCA:

$$R_{m,n,\ell} = \mathbb{E}\left\|\mathfrak{I}\tilde{k}(\cdot, X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}, \mathfrak{I}\tilde{k}(\cdot, X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\left(I - \mathfrak{U}\hat{\Sigma}_{m,\ell}^{-1}\mathfrak{U}^*\right)\mathfrak{I}\mathfrak{I}^*\right\|_{HS}^2$$

Reconstruction Error

$\left(\frac{\Im\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$. Define $\tilde{k}(\cdot,x) = k(\cdot,x) - \mu_{\mathbb{P}}$ and $\tau > 0$.

- Population KPCA:

$$R_\ell = \mathbb{E}\left\|\Im\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\Im\phi_i}{\sqrt{\lambda_i}}, \Im\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\Im\phi_i}{\sqrt{\lambda_i}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\Sigma - \Sigma^{1/2}\Sigma_\ell^{-1}\Sigma^{3/2}\right\|_{HS}^2 = \|\Sigma - \Sigma_\ell\|_{HS}^2 .$$

- Empirical KPCA:

$$R_{n,\ell} = \mathbb{E}\left\|\Im\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\Im\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}, \Im\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\Im\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\Sigma - \Sigma^{1/2}\hat{\Sigma}_\ell^{-1}\Sigma^{3/2}\right\|_{HS}^2$$

- Approximate Empirical KPCA:

$$R_{m,n,\ell} = \mathbb{E}\left\|\Im\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}, \Im\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\left(I - \mathfrak{U}\hat{\Sigma}_{m,\ell}^{-1}\mathfrak{U}^*\right)\Im\Im^*\right\|_{HS}^2$$

Reconstruction Error

$\left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$. Define $\tilde{k}(\cdot,x) = k(\cdot,x) - \mu_{\mathbb{P}}$ and $\tau > 0$.

- Population KPCA:

$$R_\ell = \mathbb{E}\left\|\mathfrak{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}, \mathfrak{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\Sigma - \Sigma^{1/2}\Sigma_\ell^{-1}\Sigma^{3/2}\right\|_{HS}^2 = \|\Sigma - \Sigma_\ell\|_{HS}^2 .$$

- Empirical KPCA:

$$R_{n,\ell} = \mathbb{E}\left\|\mathfrak{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}, \mathfrak{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\Sigma - \Sigma^{1/2}\hat{\Sigma}_\ell^{-1}\Sigma^{3/2}\right\|_{HS}^2$$

- Approximate Empirical KPCA:

$$R_{m,n,\ell} = \mathbb{E}\left\|\mathfrak{I}\tilde{k}(\cdot,X) - \sum_{i=1}^{\ell}\left\langle \frac{\mathfrak{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}, \mathfrak{I}\tilde{k}(\cdot,X)\right\rangle_{L^2(\mathbb{P})} \frac{\mathfrak{U}\hat{\phi}_{i,m}}{\sqrt{\hat{\lambda}_{i,m}}}\right\|_{L^2(\mathbb{P})}^2$$

$$= \left\|\left(I - \mathfrak{U}\hat{\Sigma}_{m,\ell}^{-1}\mathfrak{U}^*\right)\mathfrak{I}\mathfrak{I}^*\right\|_{HS}^2$$

Result

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $R_\ell \lesssim n^{-2\theta(1-\frac{1}{2\alpha})}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \quad (\textit{bias dominates}) \\ n^{-(\frac{1}{2}-\theta)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \quad (\textit{variance dominates}) \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta(1-\frac{1}{2\alpha})}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \\ n^{-(\frac{1}{2}-\theta)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \quad \text{(bias dominates)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \quad \text{(variance dominates)} \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \quad (\textit{bias dominates}) \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \quad (\textit{variance dominates}) \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \le \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \le \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \qquad (\textit{bias dominates}) \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \qquad (\textit{variance dominates}) \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Result

Clearly $R_\ell \to 0$ as $\ell \to \infty$. The goal is to study the convergence rates for R_ℓ, $R_{n,\ell}$ and $R_{m,n,\ell}$ as $\ell, m, n \to \infty$.

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^\gamma$ where $\theta > 0$ and $0 < \gamma < 1$.

- $R_\ell \lesssim n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}$
- $$R_{n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \qquad (\textit{bias dominates}) \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \qquad (\textit{variance dominates}) \end{cases}$$
- $$R_{m,n,\ell} \lesssim \begin{cases} n^{-2\theta\left(1-\frac{1}{2\alpha}\right)}, & 0 < \theta \leq \frac{\alpha}{2(3\alpha-1)} \\ n^{-\left(\frac{1}{2}-\theta\right)}, & \frac{\alpha}{2(3\alpha-1)} \leq \theta < \frac{1}{2} \end{cases}$$

  for $\gamma > 2\theta$.

No statistical loss

Convergence of Projection Operators-I

Since $\left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$, we consider

- Empirical KPCA:

$$S_{n,\ell} = \left\| \sum_{i=1}^{\ell} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} - \sum_{i=1}^{\ell} \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}} \otimes \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}} \right\|_{\text{op}}$$

- Approximate Empirical KPCA:

$$S_{m,n,\ell} = \left\| \sum_{i=1}^{\ell} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} - \sum_{i=1}^{\ell} \frac{\mathfrak{U}\hat{\phi}_{m,i}}{\sqrt{\hat{\lambda}_{m,i}}} \otimes \frac{\mathfrak{U}\hat{\phi}_{m,i}}{\sqrt{\hat{\lambda}_{m,i}}} \right\|_{\text{op}}$$

as $\ell, m, n \to \infty$.

Convergence of Projection Operators-I

Since $\left(\frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}}\right)_{i=1}^{\infty}$ form an ONS for $L^2(\mathbb{P})$, we consider

- Empirical KPCA:

$$S_{n,\ell} = \left\| \sum_{i=1}^{\ell} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} - \sum_{i=1}^{\ell} \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}} \otimes \frac{\mathfrak{I}\hat{\phi}_i}{\sqrt{\hat{\lambda}_i}} \right\|_{\mathrm{op}}$$

- Approximate Empirical KPCA:

$$S_{m,n,\ell} = \left\| \sum_{i=1}^{\ell} \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} \otimes \frac{\mathfrak{I}\phi_i}{\sqrt{\lambda_i}} - \sum_{i=1}^{\ell} \frac{\mathfrak{U}\hat{\phi}_{m,i}}{\sqrt{\hat{\lambda}_{m,i}}} \otimes \frac{\mathfrak{U}\hat{\phi}_{m,i}}{\sqrt{\hat{\lambda}_{m,i}}} \right\|_{\mathrm{op}}$$

as $\ell, m, n \to \infty$.

Convergence of Projection Operators-I

Unlike in reconstruction error, the convergence of projection operators depends on the behavior of the eigen-gap, $\delta_i = \frac{1}{2}(\lambda_i - \lambda_{i+1}), \ i \in \mathbb{N}$.

- Empirical KPCA:

$$S_{n,\ell} \lesssim \frac{1}{\delta_\ell \sqrt{n}} + \frac{1}{n^{1/4}\sqrt{\lambda_\ell}},$$

assuming $\delta_\ell \gtrsim n^{-1/2}$ and $\lambda_\ell \gtrsim n^{-1/2}$.

- Approximate Empirical KPCA:

$$S_{m,n,\ell} \lesssim \frac{1}{\delta_\ell \sqrt{m}} + \frac{1}{n^{1/4}\sqrt{\lambda_\ell}},$$

assuming $\delta_\ell \gtrsim m^{-1/2}$ and $\lambda_\ell \gtrsim m^{-1/2}$.

Convergence of Projection Operators-I

Unlike in reconstruction error, the convergence of projection operators depends on the behavior of the eigen-gap, $\delta_i = \frac{1}{2}(\lambda_i - \lambda_{i+1}), \, i \in \mathbb{N}$.

- Empirical KPCA:
$$S_{n,\ell} \lesssim \frac{1}{\delta_\ell \sqrt{n}} + \frac{1}{n^{1/4}\sqrt{\lambda_\ell}},$$
assuming $\delta_\ell \gtrsim n^{-1/2}$ and $\lambda_\ell \gtrsim n^{-1/2}$.

- Approximate Empirical KPCA:
$$S_{m,n,\ell} \lesssim \frac{1}{\delta_\ell \sqrt{m}} + \frac{1}{n^{1/4}\sqrt{\lambda_\ell}},$$
assuming $\delta_\ell \gtrsim m^{-1/2}$ and $\lambda_\ell \gtrsim m^{-1/2}$.

Convergence of Projection Operators-I

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\delta_i \gtrsim i^{-\beta}$, $\beta \geq \alpha$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^{\gamma}$ where $0 < \theta < \frac{1}{2}$ and $0 < \gamma < 1$.

- $$S_{n,\ell} \lesssim \begin{cases} n^{-\left(\frac{1}{4}-\frac{\theta}{2}\right)}, & 0 < \theta < \frac{\alpha}{2(2\beta-\alpha)} \\ n^{-\left(\frac{1}{2}-\frac{\theta\beta}{\alpha}\right)}, & \frac{\alpha}{2(2\beta-\alpha)} \leq \theta < \frac{\alpha}{2\beta} \end{cases}$$

- $$S_{m,n,\ell} \lesssim \begin{cases} n^{-\left(\frac{1}{4}-\frac{\theta}{2}\right)}, & 0 < \theta < \frac{\alpha}{2(2\beta-\alpha)}, \gamma \geq \frac{1}{2} + \theta\left(\frac{2\beta}{\alpha}-1\right) \\ n^{-\left(\frac{\gamma}{2}-\frac{\theta\beta}{\alpha}\right)}, & 0 < \theta < \frac{\alpha}{2(2\beta-\alpha)}, \frac{2\theta\beta}{\alpha} < \gamma < \frac{1}{2} + \theta\left(\frac{2\beta}{\alpha}-1\right) \end{cases}$$

Convergence of Projection Operators-I

Suppose $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\delta_i \gtrsim i^{-\beta}$, $\beta \geq \alpha$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^{\gamma}$ where $0 < \theta < \frac{1}{2}$ and $0 < \gamma < 1$.

- $$S_{n,\ell} \lesssim \begin{cases} n^{-\left(\frac{1}{4}-\frac{\theta}{2}\right)}, & 0 < \theta < \frac{\alpha}{2(2\beta-\alpha)} \\ n^{-\left(\frac{1}{2}-\frac{\theta\beta}{\alpha}\right)}, & \frac{\alpha}{2(2\beta-\alpha)} \leq \theta < \frac{\alpha}{2\beta} \end{cases}$$

- $$S_{m,n,\ell} \lesssim \begin{cases} n^{-\left(\frac{1}{4}-\frac{\theta}{2}\right)}, & 0 < \theta < \frac{\alpha}{2(2\beta-\alpha)},\ \gamma \geq \frac{1}{2} + \theta\left(\frac{2\beta}{\alpha} - 1\right) \\ n^{-\left(\frac{\gamma}{2}-\frac{\theta\beta}{\alpha}\right)}, & 0 < \theta < \frac{\alpha}{2(2\beta-\alpha)},\ \frac{2\theta\beta}{\alpha} < \gamma < \frac{1}{2} + \theta\left(\frac{2\beta}{\alpha} - 1\right) \end{cases}$$

Convergence of Projection Operators-II

- Empirical KPCA:

$$
\begin{aligned}
T_{n,\ell} &= \left\| \sum_{i=1}^{\ell} \mathfrak{I}\phi_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\phi_i - \sum_{i=1}^{\ell} \mathfrak{I}\hat{\phi}_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\hat{\phi}_i \right\|_{HS} \\
&= \left\| \Sigma^{1/2} \left(\sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i - \sum_{i=1}^{\ell} \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i \right) \Sigma^{1/2} \right\|_{HS} \\
&\lesssim \frac{\ell \lambda_\ell}{\delta_\ell \sqrt{n}}
\end{aligned}
$$

- Approximate Empirical KPCA:

$$
\begin{aligned}
T_{m,n,\ell} &= \left\| \sum_{i=1}^{\ell} \mathfrak{I}\phi_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\phi_i - \sum_{i=1}^{\ell} \mathfrak{U}\hat{\phi}_{i,m} \otimes_{L^2(\mathbb{P})} \mathfrak{U}\hat{\phi}_{i,m} \right\|_{HS} \\
&\lesssim \frac{\ell \lambda_\ell}{\delta_\ell \sqrt{n}} + \frac{1}{\sqrt{m}}
\end{aligned}
$$

- Convergence rates can be derived under the assumption $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\delta_i \gtrsim i^{-\beta}$, $\beta \geq \alpha$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^{\gamma}$ where $0 < \theta < \frac{1}{2}$ and $0 < \gamma < 1$.

Convergence of Projection Operators-II

- Empirical KPCA:

$$\begin{aligned} T_{n,\ell} &= \left\| \sum_{i=1}^{\ell} \mathfrak{I}\phi_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\phi_i - \sum_{i=1}^{\ell} \mathfrak{I}\hat{\phi}_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\hat{\phi}_i \right\|_{HS} \\ &= \left\| \Sigma^{1/2} \left(\sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i - \sum_{i=1}^{\ell} \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i \right) \Sigma^{1/2} \right\|_{HS} \\ &\lesssim \frac{\ell \lambda_\ell}{\delta_\ell \sqrt{n}} \end{aligned}$$

- Approximate Empirical KPCA:

$$\begin{aligned} T_{m,n,\ell} &= \left\| \sum_{i=1}^{\ell} \mathfrak{I}\phi_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\phi_i - \sum_{i=1}^{\ell} \mathfrak{U}\hat{\phi}_{i,m} \otimes_{L^2(\mathbb{P})} \mathfrak{U}\hat{\phi}_{i,m} \right\|_{HS} \\ &\lesssim \frac{\ell \lambda_\ell}{\delta_\ell \sqrt{n}} + \frac{1}{\sqrt{m}} \end{aligned}$$

- Convergence rates can be derived under the assumption $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\delta_i \gtrsim i^{-\beta}$, $\beta \geq \alpha$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^{\gamma}$ where $0 < \theta < \frac{1}{2}$ and $0 < \gamma < 1$.

Convergence of Projection Operators-II

- Empirical KPCA:

$$
\begin{aligned}
T_{n,\ell} &= \left\| \sum_{i=1}^{\ell} \mathfrak{I}\phi_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\phi_i - \sum_{i=1}^{\ell} \mathfrak{I}\hat{\phi}_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\hat{\phi}_i \right\|_{HS} \\
&= \left\| \Sigma^{1/2} \left(\sum_{i=1}^{\ell} \phi_i \otimes_{\mathcal{H}} \phi_i - \sum_{i=1}^{\ell} \hat{\phi}_i \otimes_{\mathcal{H}} \hat{\phi}_i \right) \Sigma^{1/2} \right\|_{HS} \\
&\lesssim \frac{\ell \lambda_\ell}{\delta_\ell \sqrt{n}}
\end{aligned}
$$

- Approximate Empirical KPCA:

$$
\begin{aligned}
T_{m,n,\ell} &= \left\| \sum_{i=1}^{\ell} \mathfrak{I}\phi_i \otimes_{L^2(\mathbb{P})} \mathfrak{I}\phi_i - \sum_{i=1}^{\ell} \mathfrak{U}\hat{\phi}_{i,m} \otimes_{L^2(\mathbb{P})} \mathfrak{U}\hat{\phi}_{i,m} \right\|_{HS} \\
&\lesssim \frac{\ell \lambda_\ell}{\delta_\ell \sqrt{n}} + \frac{1}{\sqrt{m}}
\end{aligned}
$$

- Convergence rates can be derived under the assumption $\lambda_i \asymp i^{-\alpha}$, $\alpha > \frac{1}{2}$, $\delta_i \gtrsim i^{-\beta}$, $\beta \geq \alpha$, $\ell = n^{\frac{\theta}{\alpha}}$ and $m = n^{\gamma}$ where $0 < \theta < \frac{1}{2}$ and $0 < \gamma < 1$.

Summary

- Random feature approximation to kernel PCA improves its computational complexity.
- Statistical trade-off:
  - Reconstruction error
  - Convergence of eigenspaces
- Open questions:
  - Lower bounds
  - Extension to kernel canonical correlation analysis
  - Nyström approximation

Thank You

References I

Aronszajn, N. (1950).
Theory of reproducing kernels.
Trans. Amer. Math. Soc., 68:337–404.

Fine, S. and Scheinberg, K. (2001).
Efficient SVM training using low-rank kernel representations.
Journal of Machine Learning Research, 2:243–264.

Rahimi, A. and Recht, B. (2008a).
Random features for large-scale kernel machines.
In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc.

Rahimi, A. and Recht, B. (2008b).
Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning.
In *NIPS*, pages 1313–1320.

Rudi, A. and Rosasco, L. (2016).
Generalization properties of learning with random features.
https://arxiv.org/pdf/1602.04474.pdf.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998).
Nonlinear component analysis as a kernel eigenvalue problem.
Neural Computation, 10:1299–1319.

Smola, A. J. and Schölkopf, B. (2000).
Sparse greedy matrix approximation for machine learning.
In *Proc. 17th International Conference on Machine Learning*, pages 911–918. Morgan Kaufmann, San Francisco, CA.

Sriperumbudur, B. K. and Szabó, Z. (2015).
Optimal rates for random Fourier features.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1144–1152. Curran Associates, Inc.

Sutherland, D. and Schneider, J. (2015).
On the error of random fourier features.
In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 862–871.

References II

Williams, C. and Seeger, M. (2001).
Using the Nyström method to speed up kernel machines.
In T. K. Leen, T. G. Diettrich, V. T., editor, *Advances in Neural Information Processing Systems 13*, pages 682–688, Cambridge, MA. MIT Press.

Yang, Y., Pilanci, M., and Wainwright, M. J. (2015).
Randomized sketches for kernels: Fast and optimal non-parametric regression.
Technical report.
https://arxiv.org/pdf/1501.06195.pdf.