Shrinkage Estimation in Reproducing Kernel Hilbert Space

Bharath K. Sriperumbudur

Department of Statistics, Pennsylvania State University

École Polytechnique
CMAP Seminar, July 4, 2017

Collaborators

- Prof. Kenji Fukumizu : The Institute for Statistical Mathematics, Tokyo, Japan.
- Dr. Arthur Gretton : Gatsby Computational Neuroscience Unit, University College London.
- Prof. Krikamol Muandet : Mahidol University, Bangkok, Thailand.
- Prof. Dr. Bernhard Schölkopf : Max Planck Institute for Intelligent Systems, Tübingen, Germany.

Motivating Example: Coin Toss

- Toss 1: $T\ H\ H\ H\ T\ T\ H\ T\ T\ H\ H\ T\ H$
- Toss 2: $H\ T\ T\ H\ T\ H\ T\ T\ H\ H\ H\ T\ T$

Are the coins/tosses statistically similar?

Toss 1 is a sample from $\mathbb{P}$:=Bernoulli(p) and Toss 2 is a sample from $\mathbb{Q}$:=Bernoulli(q).

Is $p = q$ or not?, i.e., compare

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\{0,1\}} x\, d\mathbb{P}(x) \qquad \text{and} \qquad \mathbb{E}_{\mathbb{Q}}[X] = \int_{\{0,1\}} x\, d\mathbb{Q}(x).$$

Motivating Example: Coin Toss

- Toss 1: $T\ H\ H\ H\ T\ T\ H\ T\ T\ H\ H\ T\ H$
- Toss 2: $H\ T\ T\ H\ T\ H\ T\ T\ H\ H\ H\ T\ T$

Are the coins/tosses statistically similar?

Toss 1 is a sample from $\mathbb{P}$:=Bernoulli(p) and Toss 2 is a sample from $\mathbb{Q}$:=Bernoulli(q).

Is $p = q$ or not?, i.e., compare

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\{0,1\}} x\, d\mathbb{P}(x) \qquad \text{and} \qquad \mathbb{E}_{\mathbb{Q}}[X] = \int_{\{0,1\}} x\, d\mathbb{Q}(x).$$

Coin Toss Example

In other words, we compare

$$\int_{\mathbb{R}} \Phi(x)\, d\mathbb{P}(x) \quad \text{and} \quad \int_{\mathbb{R}} \Phi(x)\, d\mathbb{Q}(x)$$

where Φ is an identity map,

$$\Phi(x) = x.$$

A positive definite kernel corresponding to Φ is

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathbb{R}} = xy,$$

which is a linear kernel on $\{0, 1\}$. Therefore, comparing two Bernoulli is equivalent to

$$\int_{\{0,1\}} k(y, x)\, d\mathbb{P}(x) \stackrel{?}{=} \int_{\{0,1\}} k(y, x)\, d\mathbb{Q}(x)$$

for all $y \in \{0, 1\}$, i.e., compare the expectations of the kernel.

Comparing two Gaussians

$$\mathbb{P} = N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad \mathbb{Q} = N(\mu_2, \sigma_2^2)$$

Comparing $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to comparing μ_1, μ_2 and σ_1^2, σ_2^2, i.e.,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X]$$

and

$$\mathbb{E}_{\mathbb{P}}[X^2] = \int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x^2 \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X^2].$$

Concisely

$$\int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} \Phi(x) \, d\mathbb{Q}(x)$$

where

$$\Phi(x) = (x, x^2).$$

Compare the first moment of the feature map

Comparing two Gaussians

$$\mathbb{P} = N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad \mathbb{Q} = N(\mu_2, \sigma_2^2)$$

Comparing $\mathbb{P}$ and $\mathbb{Q}$ is equivalent to comparing μ_1, μ_2 and σ_1^2, σ_2^2, i.e.,

$$\mathbb{E}_{\mathbb{P}}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X]$$

and

$$\mathbb{E}_{\mathbb{P}}[X^2] = \int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} x^2 \, d\mathbb{Q}(x) = \mathbb{E}_{\mathbb{Q}}[X^2].$$

Concisely

$$\int_{\mathbb{R}} \Phi(x) \, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} \Phi(x) \, d\mathbb{Q}(x)$$

where

$$\Phi(x) = (x, x^2).$$

Compare the first moment of the feature map

Comparing two Gaussians

Using the map Φ, we can construct a positive definite kernel as

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathbb{R}^2} = xy + x^2 y^2$$

which is a polynomial kernel of order 2.

Therefore, comparing two Gaussians is equivalent to

$$\int_{\mathbb{R}} k(y, x)\, d\mathbb{P}(x) \stackrel{?}{=} \int_{\mathbb{R}} k(y, x)\, d\mathbb{Q}(x)$$

for all $y \in \mathbb{R}$, i.e., compare the expectations of the kernel.

Comparing general $\mathbb{P}$ and $\mathbb{Q}$

Moment generating function is defined as

$$M_{\mathbb{P}}(y) = \int_{\mathbb{R}} e^{xy}\, d\mathbb{P}(x)$$

and (if it exists) captures the information about a distribution, i.e.,

$$M_{\mathbb{P}} = M_{\mathbb{Q}} \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Choosing

$$\Phi(x) = \left(1, x, \frac{x^2}{\sqrt{2!}}, \ldots, \frac{x^i}{\sqrt{i!}}, \ldots\right) \in \ell_2(\mathbb{N}),\ \forall\, x \in \mathbb{R}$$

it is easy to verify that

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\ell_2(\mathbb{N})} = e^{xy}$$

and so

$$\int_{\mathbb{R}} k(x, y)\, d\mathbb{P}(x) = \int_{\mathbb{R}} k(x, y)\, d\mathbb{Q}(x),\ \forall\, y \in \mathbb{R} \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Two-Sample Problem

- Given random samples $\{X_1, \ldots, X_m\} \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_1, \ldots, Y_n\} \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: $\mathbb{P} = \mathbb{Q}$ or $\mathbb{P} \neq \mathbb{Q}$?

Applications:

- Microarray data (aggregation problem)
- Speaker verification
- Independence Testing: Given random samples $\{(X_1, Y_1), \ldots, (X_n, Y_n)\} \overset{i.i.d.}{\sim} \mathbb{P}_{xy}$. Does $\mathbb{P}_{xy}$ factorize into $\mathbb{P}_x \mathbb{P}_y$?
- Feature selection (microarrays, image and text,...)

Kernel Mean

- Canonical feature map:

$$\Phi(x) = k(\cdot, x) \in \mathcal{H}, \qquad x \in \mathcal{X}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS).

- Generalization to probabilities:

$$x \mapsto k(\cdot, x) \qquad \equiv \qquad \underbrace{\delta_x}_{\text{point mass at } x} \qquad \mapsto \qquad \underbrace{k(\cdot, x)}_{\int_{\mathcal{X}} k(\cdot, y)\, d\delta_x(y)}$$

Based on the above, the map is extended to probability measures as

$$\mathbb{P} \mapsto \Phi(\mathbb{P}) := \underbrace{\int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)}_{\mathbb{E}_{X \sim \mathbb{P}} k(\cdot, X)}$$

(Smola et al., 2007)

Why is this useful?

- For an appropriate choice of k, $\Phi(\mathbb{P})$ can distinguish $\mathbb{P}$ by high-order moments.

$$k(y,x)=c_0 + c_1(xy) + c_2(xy)^2 + \cdots \ (c_i \neq 0)$$
$$\Phi(\mathbb{P})(y)=c_0 + c_1 \left(\int_{\mathbb{R}} x \, d\mathbb{P}(x) \right) y + c_2 \left(\int_{\mathbb{R}} x^2 \, d\mathbb{P}(x) \right) y^2 + \cdots$$

The kernel mean is a generalization of

- Characteristic function: $k(\cdot, x) = e^{\sqrt{-1}\langle \cdot, x \rangle}, \ x, y \in \mathbb{R}^d$
- Moment generating function: $k(\cdot, x) = e^{\langle \cdot, x \rangle}, \ x, y \in \mathbb{R}^d$
- Weierstrass transform: $k(\cdot, x) = (4\pi)^{-d/2} e^{-\|\cdot - x\|^2/4}$

to *arbitrary* space $\mathcal{X}$.

Many Applications

- Two-sample testing (Gretton et al., 2007)
- Independence testing (Gretton et al., 2008)
- Feature selection (Song et al., 2012)
- Kernel Bayes' rule (Fukumizu et al., 2013)
- Density estimation (S, 2011)
- Causal inference (Lopez-Paz et al., 2015)
- Distribution regression (Szabó et al., 2015), ...

Reproducing Kernel Hilbert Space

A Hilbert space, $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS) if the evaluational functional

$$\delta_x : \mathcal{H} \to \mathbb{R}, \qquad f \mapsto f(x)$$

is continuous for each $x \in \mathcal{X}$.

- Riesz representation: $\forall\, x \in \mathcal{X}$, $\exists$ unique $k_x \in \mathcal{H}$ such that

$$\delta_x(f) = f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \ \ \forall\, f \in \mathcal{H}.$$

- Reproducing property, symmetry and positive-definiteness:

$$k(y, x) := k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = k_x(y) = k(x, y), \ \ x, y \in \mathcal{X}.$$

  $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel.

- Moore-Aronszajn Theorem: For every positive definite function, k on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS, $\mathcal{H}$ with k as its r.k. (Aronszajn-50)

Reproducing Kernel Hilbert Space

A Hilbert space, $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS) if the evaluational functional

$$\delta_x : \mathcal{H} \to \mathbb{R}, \qquad f \mapsto f(x)$$

is continuous for each $x \in \mathcal{X}$.

- Riesz representation: $\forall\, x \in \mathcal{X}$, $\exists$ unique $k_x \in \mathcal{H}$ such that

$$\delta_x(f) = f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \ \ \forall\, f \in \mathcal{H}.$$

- Reproducing property, symmetry and positive-definiteness:

$$k(y, x) := k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = k_x(y) = k(x, y), \ \ x, y \in \mathcal{X}.$$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel.

- Moore-Aronszajn Theorem: For every positive definite function, k on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS, $\mathcal{H}$ with k as its r.k. (Aronszajn-50)

Reproducing Kernel Hilbert Space

A Hilbert space, $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS) if the evaluational functional

$$\delta_x : \mathcal{H} \to \mathbb{R}, \qquad f \mapsto f(x)$$

is continuous for each $x \in \mathcal{X}$.

- Riesz representation: $\forall\, x \in \mathcal{X}$, $\exists$ unique $k_x \in \mathcal{H}$ such that

$$\delta_x(f) = f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \ \ \forall\, f \in \mathcal{H}.$$

- Reproducing property, symmetry and positive-definiteness:

$$k(y, x) := k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = k_x(y) = k(x, y), \ \ x, y \in \mathcal{X}.$$

  $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel.

- Moore-Aronszajn Theorem: For every positive definite function, k on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS, $\mathcal{H}$ with k as its r.k. (Aronszajn-50)

Reproducing Kernel Hilbert Space

A Hilbert space, $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is called a reproducing kernel Hilbert space (RKHS) if the evaluational functional

$$\delta_x : \mathcal{H} \to \mathbb{R}, \qquad f \mapsto f(x)$$

is continuous for each $x \in \mathcal{X}$.

- Riesz representation: $\forall\, x \in \mathcal{X}$, $\exists$ unique $k_x \in \mathcal{H}$ such that

$$\delta_x(f) = f(x) = \langle f, k_x \rangle_{\mathcal{H}}, \ \ \forall\, f \in \mathcal{H}.$$

- Reproducing property, symmetry and positive-definiteness:

$$k(y, x) := k_y(x) = \langle k_y, k_x \rangle_{\mathcal{H}} = k_x(y) = k(x, y), \ \ x, y \in \mathcal{X}.$$

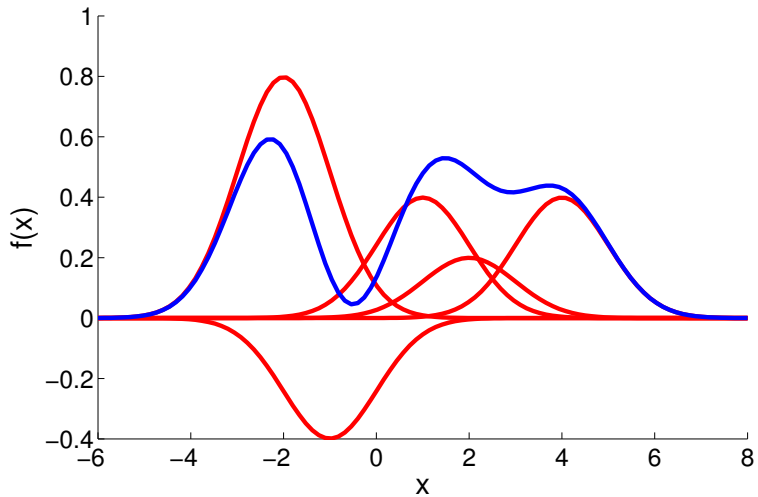
  $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel.

- Moore-Aronszajn Theorem: For every positive definite function, k on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS, $\mathcal{H}$ with k as its r.k. (Aronszajn-50)

Functions in the RKHS

- $\mathcal{H} = \overline{\text{span}\{k(\cdot, x) : x \in \mathcal{X}\}}$ (linear span of kernel functions)
  - Example: $f(x) = \sum_{i=1}^{m} \alpha_i k(x, x_i)$ for arbitrary $m \in \mathbb{N}$, $\{\alpha_i\} \subset \mathbb{R}$, $x \in \mathcal{X}$ and $\{x_i\} \subset \mathcal{X}$.

$$k(x, y) = e^{-\|x-y\|^2/2\sigma^2}$$



Picture credit: A. Gretton

Properties of RKHS

- k is bounded if and only every $f \in \mathcal{H}$ is bounded.
- If $\int_{\mathcal{X}} \sqrt{k(x,x)}\, d\mu(x) < \infty$, then for every $f \in \mathcal{H}$, $\int_{\mathcal{X}} f(x)\, d\mu(x) < \infty$.
- Every $f \in \mathcal{H}$ is continuous if and only if $k(\cdot, x)$ is continuous for all $x \in \mathcal{X}$.
- Every $f \in \mathcal{H}$ is m-times continuously differentiable if k is m-times continuously differentiable.

k controls the properties of $\mathcal{H}$

Positive Definiteness: Translation Invariant Kernels

- Positive definiteness: k is said to be positive definite if for all $n \in \mathbb{N}$, $(\alpha_i)_{i=1}^n \subset \mathbb{R}$ and all $(x_i)_{i=1}^n \subset \mathcal{X}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

- Let $\mathcal{X} = \mathbb{R}^d$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ is said to be translation invariant if

$$k(x, y) = \psi(x - y), \; x, y \in \mathbb{R}^d,$$

where ψ is a positive definite function on $\mathbb{R}^d$.

- Bochner's theorem provides a complete characterization for the positive definiteness of ψ.
- A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if ψ is the Fourier transform of a finite non-negative Borel measure Λ, i.e.,

$$\psi(x) = \underbrace{\int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x, \omega \rangle_2} \, d\Lambda(\omega)}_{\text{Characteristic function of } \Lambda}.$$

Positive Definiteness: Translation Invariant Kernels

- Positive definiteness: k is said to be positive definite if for all $n \in \mathbb{N}$, $(\alpha_i)_{i=1}^n \subset \mathbb{R}$ and all $(x_i)_{i=1}^n \subset \mathcal{X}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

- Let $\mathcal{X} = \mathbb{R}^d$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ is said to be translation invariant if

$$k(x, y) = \psi(x - y), \ x, y \in \mathbb{R}^d,$$

where ψ is a positive definite function on $\mathbb{R}^d$.

- Bochner's theorem provides a complete characterization for the positive definiteness of ψ.
- A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if ψ is the Fourier transform of a finite non-negative Borel measure Λ, i.e.,

$$\psi(x) = \underbrace{\int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x, \omega \rangle_2} \, d\Lambda(\omega)}_{\text{Characteristic function of } \Lambda}.$$

Positive Definiteness: Translation Invariant Kernels

- Positive definiteness: k is said to be positive definite if for all $n \in \mathbb{N}$, $(\alpha_i)_{i=1}^n \subset \mathbb{R}$ and all $(x_i)_{i=1}^n \subset \mathcal{X}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

- Let $\mathcal{X} = \mathbb{R}^d$. A kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^d$ is said to be translation invariant if

$$k(x, y) = \psi(x - y), \ x, y \in \mathbb{R}^d,$$

where ψ is a positive definite function on $\mathbb{R}^d$.

- Bochner's theorem provides a complete characterization for the positive definiteness of ψ.
- A continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if ψ is the Fourier transform of a finite non-negative Borel measure Λ, i.e.,

$$\psi(x) = \underbrace{\int_{\mathbb{R}^d} e^{\sqrt{-1}\langle x, \omega \rangle_2} \, d\Lambda(\omega)}_{\text{Characteristic function of } \Lambda}.$$

Explicit Realization of RKHS

- $\mathcal{X} = \mathbb{R}^d$ and $k(x, y) = \psi(x - y)$ where ψ is a positive definite function.
- Let $\psi \in L^1(\mathcal{X})$. Then

$$\mathcal{H} = \left\{ f \in L^2(\mathcal{X}) \cap C_b(\mathcal{X}) \middle| \int \frac{|\hat{f}(\omega)|^2}{\hat{\psi}(\omega)} \, d\omega < \infty \right\}$$

endowed with

$$\langle f, g \rangle_{\mathcal{H}} = (2\pi)^{-d/2} \int \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\psi}(\omega)} \, d\omega$$

is an RKHS with k as the r.k., where $\hat{\psi}$ is the Fourier transform of ψ.

(Wendland, 2005)

Gaussian RKHS

- Gaussian kernel:

$$k(x, y) = \psi(x - y) = e^{-\|x-y\|_2^2/\gamma^2},\ x, y \in \mathbb{R}^d$$

- Fourier transform:

$$\hat{\psi}(\omega) = \left(\frac{\gamma^2}{2}\right)^{d/2} e^{-\frac{\gamma^2\|\omega\|_2^2}{4}},\ \omega \in \mathbb{R}^d$$

-

$$\mathcal{H}_\gamma(\mathbb{R}^d) := \left\{ f \in L^2(\mathbb{R}^d) : \underbrace{\int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 e^{\frac{\gamma^2\|\omega\|_2^2}{4}} d\omega}_{\|f\|_{\mathcal{H}_\gamma}^2} < \infty \right\}$$

Fast decay of $\hat{\psi} \Rightarrow$ Smooth $\mathcal{H}$

Sobolev RKHS

- Laplacian kernel:

$$k(x,y) = \psi(x-y) = \sqrt{\frac{\pi}{2}} e^{-|x-y|},\ x,y \in \mathbb{R}$$

- Fourier transform:

$$\hat{\psi}(\omega) = \frac{1}{1+|\omega|^2},\ \omega \in \mathbb{R}$$

-

$$\mathcal{H}_1^2(\mathbb{R}) := \left\{ f \in L^2(\mathbb{R}) \,:\, \underbrace{\int_{\mathbb{R}} |\hat{f}(\omega)|^2 (1+|\omega|^2)\, d\omega}_{\|f\|^2_{\mathcal{H}_1^2}} < \infty \right\}$$

Extension to $\mathbb{R}^d$: Matérn Kernel

Two-Sample Problem

- Given random samples $\{X_j\}_{j=1}^m \overset{i.i.d.}{\sim} \mathbb{P}$ and $\{Y_j\}_{j=1}^n \overset{i.i.d.}{\sim} \mathbb{Q}$.
- Determine: are $\mathbb{P}$ and $\mathbb{Q}$ different?
- Approach:

$$\begin{array}{ccc} H_0 : \mathbb{P} = \mathbb{Q} & & H_0 : \rho(\mathbb{P}, \mathbb{Q}) = 0 \\ & \equiv & \\ H_1 : \mathbb{P} \neq \mathbb{Q} & & H_1 : \rho(\mathbb{P}, \mathbb{Q}) > 0 \end{array}$$

  where ρ is a distance on probabilities.

- If empirical ρ is
  - far from zero: reject H_0
  - close to zero: accept H_0

Kernel Mean

- Using the kernel mean,

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

we define

$$\rho(\mathbb{P}, \mathbb{Q}) := \left\| \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}}$$

called the maximum mean discrepancy (Gretton et al., 2007).

- For appropriate kernels (e.g., Gaussian kernel),

$$\mathbb{P} \mapsto \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x) \text{ is injective,}$$

i.e.,

$$\rho(\mathbb{P}, \mathbb{Q}) = 0 \qquad \Leftrightarrow \qquad \mathbb{P} = \mathbb{Q}.$$

Kernel Mean Estimation

Estimating ρ involves estimating the kernel means $\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$.

- Given: $(X_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.
- Goal: Estimate $\mu_{\mathbb{P}}$.

Empirical estimator of $\mu_{\mathbb{P}}$:

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i)$$

is the non-parametric maximum likelihood estimator (MLE) of $\mu_{\mathbb{P}}$.

Applications:

- The test statistic in two-sample testing is:

$$\|\hat{\mu}_{\mathbb{P}} - \hat{\mu}_{\mathbb{Q}}\|_{\mathcal{H}}^2 = \frac{1}{n^2} \sum_{i,j}^{n} k(X_i, X_j) + \frac{1}{m^2} \sum_{i,j}^{m} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i,j} k(X_i, Y_j)$$

Simple to compute!!

Theoretical Results

- $\hat{\mu}_{\mathbb{P}}$ is a $\sqrt{n}$-consistent estimator of $\mu_{\mathbb{P}}$, i.e.,

$$\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}} = O_{\mathbb{P}}\left(n^{-1/2}\right) \text{ as } n \to \infty.$$

- The minimax rate of estimating $\mu_{\mathbb{P}}$ is $n^{-1/2}$, i.e.,

$$\inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n \left(\sqrt{n}\|\hat{\theta} - \mu_{\mathbb{P}}\|_{\mathcal{H}} \geq c_{\mathcal{P}}\right) > 0$$

for some suitable $\mathcal{P}$ and constant $c_{\mathcal{P}}$ (Tolstikhin et al., 2016).

$\hat{\mu}_{\mathbb{P}}$ is a minimax rate optimal estimator of $\mu_{\mathbb{P}}$.

Question

Can we improve upon $\hat{\mu}_{\mathbb{P}}$? Particularly, can we construct $\check{\mu}_{\mathbb{P}}$ such that

$$\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2, \ \forall\, \mathbb{P} \in \mathcal{P}$$

and there exists a $\mathbb{P} \in \mathcal{P}$ for which

$$\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 ?$$

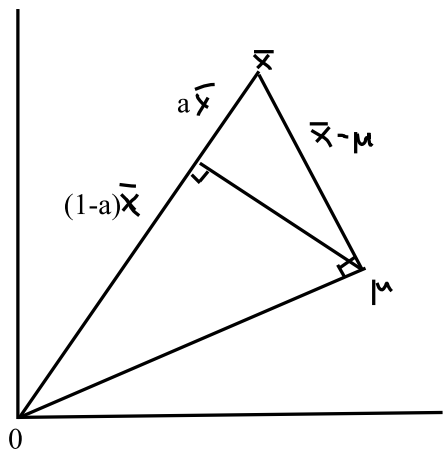
Motivation

- Setting:
  - $\mathcal{X} = \mathbb{R}^d$, $k(\cdot, x) = x$, $\mathbb{P} \sim N(\mu, \sigma^2 I_d)$ where σ^2 is known.
- Maximum likelihood estimator:

$$\hat{\mu}_{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}$$

is a sufficient statistic, the minimum variance unbiased estimator and is minimax optimal.

$$\mathbb{E}\|\overline{X}\|_2^2 = \frac{d\sigma^2}{n} + \|\mu\|_2^2 \geq \|\mu\|_2^2 \quad \text{and} \quad \mathbb{E}\langle \overline{X} - \mu, \mu \rangle_2 = 0.$$



James-Stein Shrinkage

For an appropriate choice of a, it is possible that

$$\|(1-a)\overline{X} - \mu\|_2^2 < \|\overline{X} - \mu\|_2^2, \ \ \forall \mu.$$

James-Stein Shrinkage estimator:

$$\check{\mu}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{n\|\overline{X}\|^2}\right)\overline{X}$$

improves upon $\overline{X}$ for $d \geq 3$ (James and Stein, 1961).

- Similar behavior occurs in RKHS $\mathcal{H}$ wherein for any $f^* \in \mathcal{H}$,

$$\mathbb{E}\|\hat{\mu}_{\mathbb{P}} - f^*\|_{\mathcal{H}}^2 \geq \|\mu_{\mathbb{P}} - f^*\|_{\mathcal{H}}^2$$

and so the shrinkage estimator to consider is

$$(1-\alpha)(\mu_{\mathbb{P}} - f^*) + f^*.$$

James-Stein Shrinkage

For an appropriate choice of a, it is possible that

$$\|(1-a)\overline{X} - \mu\|_2^2 < \|\overline{X} - \mu\|_2^2, \ \ \forall\, \mu.$$

James-Stein Shrinkage estimator:

$$\check{\mu}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{n\|\overline{X}\|^2}\right)\overline{X}$$

improves upon $\overline{X}$ for $d \geq 3$ (James and Stein, 1961).

- Similar behavior occurs in RKHS $\mathcal{H}$ wherein for any $f^* \in \mathcal{H}$,

$$\mathbb{E}\|\hat{\mu}_{\mathbb{P}} - f^*\|_{\mathcal{H}}^2 \geq \|\mu_{\mathbb{P}} - f^*\|_{\mathcal{H}}^2$$

and so the shrinkage estimator to consider is

$$(1-\alpha)(\mu_{\mathbb{P}} - f^*) + f^*.$$

Shrinkage Estimator of $\mu_{\mathbb{P}}$

For $\alpha \in \mathbb{R}$ and $f^* \in \mathcal{H}$, define

$$\check{\mu}_\alpha = (1-\alpha)(\hat{\mu}_{\mathbb{P}} - f^*) + f^*,$$

$$\Delta := \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \text{ and } \Delta_\alpha := \mathbb{E}\|\check{\mu}_\alpha - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}.$$

Theorem

For all distributions $\mathbb{P}$ and kernel k satisfying $\int k(x,x)\, d\mathbb{P}(x) < \infty$, $\Delta_\alpha < \Delta$ if and only if

$$\alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}\right). \qquad (\star)$$

In particular, $\Delta_\alpha - \Delta$ is minimized at $\alpha_* := \frac{\Delta}{\Delta + \|f^* - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}$.

- For any α satisfying $(\star)$, the positive-part estimator,

$$\check{\mu}^+_\alpha = \alpha f^* + (1-\alpha)_+ \hat{\mu}_{\mathbb{P}}$$

satisfies $\Delta^+_\alpha \leq \Delta_\alpha < \Delta$, where $\Delta^+_\alpha := \mathbb{E}\|\check{\mu}^+_\alpha - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}$.

Shrinkage Estimator of $\mu_{\mathbb{P}}$

For $\alpha \in \mathbb{R}$ and $f^* \in \mathcal{H}$, define

$$\check{\mu}_\alpha = (1-\alpha)(\hat{\mu}_{\mathbb{P}} - f^*) + f^*,$$

$$\Delta := \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \text{ and } \Delta_\alpha := \mathbb{E}\|\check{\mu}_\alpha - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}.$$

Theorem

For all distributions $\mathbb{P}$ and kernel k satisfying $\int k(x,x)\, d\mathbb{P}(x) < \infty$, $\Delta_\alpha < \Delta$ if and only if

$$\alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}\right). \qquad (\star)$$

In particular, $\Delta_\alpha - \Delta$ is minimized at $\alpha_* := \frac{\Delta}{\Delta + \|f^* - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}$.

- For any α satisfying $(\star)$, the positive-part estimator,

$$\check{\mu}^+_\alpha = \alpha f^* + (1-\alpha)_+ \hat{\mu}_{\mathbb{P}}$$

satisfies $\Delta^+_\alpha \leq \Delta_\alpha < \Delta$, where $\Delta^+_\alpha := \mathbb{E}\|\check{\mu}^+_\alpha - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}$.

Shrinkage Estimator of $\mu_{\mathbb{P}}$

For $\alpha \in \mathbb{R}$ and $f^* \in \mathcal{H}$, define

$$\check{\mu}_\alpha = (1-\alpha)(\hat{\mu}_{\mathbb{P}} - f^*) + f^*,$$

$$\Delta := \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \text{ and } \Delta_\alpha := \mathbb{E}\|\check{\mu}_\alpha - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}.$$

Theorem

For all distributions $\mathbb{P}$ and kernel k satisfying $\int k(x,x)\, d\mathbb{P}(x) < \infty$, $\Delta_\alpha < \Delta$ if and only if

$$\alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}\right). \qquad (\star)$$

In particular, $\Delta_\alpha - \Delta$ is minimized at $\alpha_* := \frac{\Delta}{\Delta + \|f^* - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}$.

- For any α satisfying $(\star)$, the positive-part estimator,

$$\check{\mu}^+_\alpha = \alpha f^* + (1-\alpha)_+ \hat{\mu}_{\mathbb{P}}$$

satisfies $\Delta^+_\alpha \leq \Delta_\alpha < \Delta$, where $\Delta^+_\alpha := \mathbb{E}\|\check{\mu}^+_\alpha - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}$.

Choice of α

Theorem

Let

- $k(x, y) = \psi(x - y) \neq 0$, $x, y \in \mathbb{R}^d$
- $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ *and* ψ *is a positive definite function.*

For a given constant $A \in (0, 1)$, *let* $A_\psi := \frac{A(2\pi)^{d/2}\psi(0)}{\|\psi\|_{L_1}}$ *and*

$$\mathcal{P}_{k,A} := \left\{ \mathbb{P} \in M_+^1(\mathbb{R}^d) : \|\phi_{\mathbb{P}}\|_{L^2} \leq \sqrt{A_\psi} \right\},$$

where $\phi_{\mathbb{P}}$ *denotes the characteristic function of* $\mathbb{P}$. *Then*

$$\forall\, \mathbb{P} \in \mathcal{P}_{k,A},\ \Delta_\alpha < \Delta$$

if $\alpha \in \left(0, \frac{2(1-A)}{1+(n-1)A+\frac{n\|f^*\|_{\mathcal{H}}^2}{\psi(0)}+\frac{2n\sqrt{A}\|f^*\|_{\mathcal{H}}}{\sqrt{\psi(0)}}} \right]$.

Examples

- $\mathbb{P} \sim N(\mu, \sigma^2 I_d)$ and $\psi(x) = e^{-\|x\|^2/2\tau^2}$:

$$\mathcal{P}_{k,A} = \left\{ \text{family of normals} : \sigma^2 \geq \frac{\pi\tau^2}{A^{2/d}} \right\}.$$

- $\mathbb{P} \overset{\text{moment}}{\sim} (\mu, \Sigma)$ and $k(x, y) = x^T y,\ x, y \in \mathbb{R}^d$:

$$\mathcal{P}_{k,A} = \left\{ \mathbb{P} \in M_+^1(\mathbb{R}^d) \,\middle|\, \frac{\|\mu\|_2^2}{\text{trace}(\Sigma)} \leq \frac{A}{1-A} \right\}.$$

Data-dependent Choice of α (B-KMSE)

Assuming $f^* = 0$, define $\hat{\alpha} := \frac{\hat{\Delta}}{\hat{\Delta}+\|\hat{\mu}\|^2_{\mathcal{H}}}$ so that $\breve{\mu}_{\hat{\alpha}} = (1-\hat{\alpha})\hat{\mu}_{\mathbb{P}}$, where

$$\hat{\Delta} = \frac{1}{n^2}\sum_{i=1}^{n} k(X_i, X_i) - \frac{1}{n^2(n-1)}\sum_{i\neq j}^{n} k(X_i, X_j)$$

and $\|\hat{\mu}_{\mathbb{P}}\|^2_{\mathcal{H}} = \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j)$.

Theorem (Oracle inequality)

Suppose $n \geq 2$. Under some moment conditions on continuous k,

$$|\hat{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$$

and

$$\min_{\alpha} \mathbb{E}\|\breve{\mu}_{\alpha} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \leq \underbrace{\mathbb{E}\|\breve{\mu}_{\hat{\alpha}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}_{\asymp n^{-1}} \leq \underbrace{\min_{\alpha} \mathbb{E}\|\breve{\mu}_{\alpha} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}_{< \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \asymp n^{-1}} + O(n^{-3/2})$$

as $n \to \infty$.

Data-dependent Choice of α (B-KMSE)

Assuming $f^* = 0$, define $\hat{\alpha} := \frac{\hat{\Delta}}{\hat{\Delta}+\|\hat{\mu}\|^2_{\mathcal{H}}}$ so that $\breve{\mu}_{\hat{\alpha}} = (1-\hat{\alpha})\hat{\mu}_{\mathbb{P}}$, where

$$\hat{\Delta} = \frac{1}{n^2}\sum_{i=1}^{n} k(X_i, X_i) - \frac{1}{n^2(n-1)}\sum_{i\neq j}^{n} k(X_i, X_j)$$

and $\|\hat{\mu}_{\mathbb{P}}\|^2_{\mathcal{H}} = \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j)$.

Theorem (Oracle inequality)

Suppose $n \geq 2$. Under some moment conditions on continuous k,

$$|\hat{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$$

and

$$\min_{\alpha} \mathbb{E}\|\breve{\mu}_{\alpha} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \leq \underbrace{\mathbb{E}\|\breve{\mu}_{\hat{\alpha}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}_{\asymp n^{-1}} \leq \underbrace{\min_{\alpha} \mathbb{E}\|\breve{\mu}_{\alpha} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}}}_{< \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|^2_{\mathcal{H}} \asymp n^{-1}} + O(n^{-3/2})$$

as $n \to \infty$.

Relation to Stein Shrinkage I

Suppose $\mathbb{P} \sim N(\mu, \sigma^2 I)$ and $k(x, y) = x^T y$. Then

$$\check{\mu}_{\hat{\alpha}} = \frac{n\|\overline{X}\|^2}{S^2 + n\|\overline{X}\|^2}\overline{X} = \left(1 - \frac{S^2}{S^2 + n\|\overline{X}\|^2}\right)\overline{X},$$

whereas

$$\hat{\mu}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{n\|\overline{X}\|^2}\right)\overline{X}$$

is the James-Stein shrinkage estimator.

- For $d \geq 4 + \frac{2}{n-1}$,

$$\mathbb{E}\|\check{\mu}_{\hat{\alpha}} - \mu\|^2 < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu\|^2$$

  for all $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$.

- Weaker than Stein's result but the estimator is constructed without requiring the knowledge of $\mathbb{P}$.

Relation to Stein Shrinkage I

Suppose $\mathbb{P} \sim N(\mu, \sigma^2 I)$ and $k(x, y) = x^T y$. Then

$$\check{\mu}_{\hat{\alpha}} = \frac{n\|\overline{X}\|^2}{S^2 + n\|\overline{X}\|^2}\overline{X} = \left(1 - \frac{S^2}{S^2 + n\|\overline{X}\|^2}\right)\overline{X},$$

whereas

$$\hat{\mu}_{JS} = \left(1 - \frac{(d-2)\sigma^2}{n\|\overline{X}\|^2}\right)\overline{X}$$

is the James-Stein shrinkage estimator.

- For $d \geq 4 + \frac{2}{n-1}$,

$$\mathbb{E}\|\check{\mu}_{\hat{\alpha}} - \mu\|^2 < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu\|^2$$

  for all $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$.

- Weaker than Stein's result but the estimator is constructed without requiring the knowledge of $\mathbb{P}$.

Relation to Stein Shrinkage II

Suppose we modify the estimator $\check{\mu}_{\hat{\alpha}}$ as

$$\check{\mu}_{\hat{\alpha}} = \left(1 - c\frac{S^2}{S^2 + n\|\overline{X}\|^2}\right)\overline{X}.$$

- If

$$c = \frac{2n-2}{3n-1},$$

then for $d \geq 3$,

$$\mathbb{E}\|\check{\mu}_{\hat{\alpha}} - \mu\|^2 < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu\|^2$$

for all $\mu \in \mathbb{R}^d$ and $\sigma^2 > 0$.

Shrinkage and Regularization

- Observation:

$$\mu_{\mathbb{P}} = \arg \inf_{g \in \mathcal{H}} \int \|k(\cdot, x) - g\|_{\mathcal{H}}^2 \, d\mathbb{P}(x)$$

and

$$\hat{\mu}_{\mathbb{P}} = \arg \inf_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|k(\cdot, X_i) - g\|_{\mathcal{H}}^2.$$

- No need for regularization: What if?

$$\check{\mu}_{\lambda} = \arg \inf_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|k(\cdot, X_i) - g\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{H}}^2 = \frac{\hat{\mu}_{\mathbb{P}}}{1 + \lambda},$$

equivalent to shrinkage estimator with $\alpha = \frac{\lambda}{1+\lambda}$ and $f^* = 0$, where $\lambda > 0$.

Shrinkage and Regularization

- Observation:

$$\mu_{\mathbb{P}} = \arg \inf_{g \in \mathcal{H}} \int \|k(\cdot, x) - g\|_{\mathcal{H}}^2 \, d\mathbb{P}(x)$$

and

$$\hat{\mu}_{\mathbb{P}} = \arg \inf_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|k(\cdot, X_i) - g\|_{\mathcal{H}}^2.$$

- No need for regularization: What if?

$$\check{\mu}_{\lambda} = \arg \inf_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|k(\cdot, X_i) - g\|_{\mathcal{H}}^2 + \lambda \|g\|_{\mathcal{H}}^2 = \frac{\hat{\mu}_{\mathbb{P}}}{1 + \lambda},$$

equivalent to shrinkage estimator with $\alpha = \frac{\lambda}{1+\lambda}$ and $f^* = 0$, where $\lambda > 0$.

Choice of λ (R-KMSE)

Let $n \geq 2$, $\rho := \frac{1}{n^2}\sum_{i,j=1}^{n} k(X_i, X_j)$ and $\theta := \frac{1}{n}\sum_{i=1}^{n} k(X_i, X_i)$. Assuming $n\rho > \theta$, the unique minimizer of

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left\| k(\cdot, X_i) - \breve{\mu}_{\lambda}^{(-i)} \right\|_{\mathcal{H}}^2$$

is given by

$$\hat{\lambda} = \frac{n(\theta - \rho)}{(n-1)(n\rho - \theta)}$$

and is closely related to B-KMSE and is a positive-part estimator.

Oracle Inequality for R-KMSE

$$\min_{\alpha} \mathbb{E}\|\check{\mu}_{\alpha} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\check{\mu}_{\hat{\lambda}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 \leq \min_{\alpha} \mathbb{E}\|\check{\mu}_{\alpha} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 + O(n^{-3/2})$$

where $\check{\mu}_{\alpha} = (1-\alpha)\hat{\mu}_{\mathbb{P}}$ and therefore

$$\underbrace{\mathbb{E}\|\check{\mu}_{\hat{\lambda}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2}_{\asymp n^{-1}} < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 + O(n^{-3/2}).$$

as $n \to \infty$.

Spectral Shrinkage (S-KMSE)

In the regularization interpretation of shrinkage, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\|k(\cdot,X_i)-g\|_{\mathcal{H}}^2+\lambda\|g\|_{\mathcal{H}}^2$$

is of the form

$$g=\sum_{i=1}^{n}\beta_i k(\cdot,X_i). \qquad (\star\star)$$

Suppose we are interested in estimators of the form $(\star\star)$ where β is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\|k(\cdot,X_i)-\sum_{j=1}^{n}\beta_j k(\cdot,X_i)\|_{\mathcal{H}}^2+\lambda\|\beta\|_2^2.$$

It can be shown that $\beta=(\mathbf{K}+n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n$ resulting in an estimator,

$$\check{\mu}_\lambda=\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\hat{\mu}_{\mathbb{P}}=\sum_{i=1}^{n}\frac{\gamma_i}{\gamma_i+\lambda}\langle\hat{\mu}_{\mathbb{P}},\phi_i\rangle_{\mathcal{H}}\phi_i,$$

where $\hat{\Sigma}$ is the empirical covariance operator on $\mathcal{H}$

Spectral Shrinkage (S-KMSE)

In the regularization interpretation of shrinkage, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\|k(\cdot,X_i)-g\|_{\mathcal{H}}^2+\lambda\|g\|_{\mathcal{H}}^2$$

is of the form

$$g=\sum_{i=1}^{n}\beta_i k(\cdot,X_i). \qquad (\star\star)$$

Suppose we are interested in estimators of the form $(\star\star)$ where $\boldsymbol{\beta}$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\|k(\cdot,X_i)-\sum_{j=1}^{n}\beta_j k(\cdot,X_i)\|_{\mathcal{H}}^2+\lambda\|\boldsymbol{\beta}\|_2^2.$$

It can be shown that $\boldsymbol{\beta}=(\mathbf{K}+n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n$ resulting in an estimator,

$$\check{\mu}_\lambda=\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\hat{\mu}_{\mathbb{P}}=\sum_{i=1}^{n}\frac{\gamma_i}{\gamma_i+\lambda}\langle\hat{\mu}_{\mathbb{P}},\phi_i\rangle_{\mathcal{H}}\phi_i,$$

where $\hat{\Sigma}$ is the empirical covariance operator on $\mathcal{H}$

Spectral Shrinkage (S-KMSE)

In the regularization interpretation of shrinkage, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\|k(\cdot,X_i)-g\|_{\mathcal{H}}^2+\lambda\|g\|_{\mathcal{H}}^2$$

is of the form

$$g=\sum_{i=1}^{n}\beta_i k(\cdot,X_i). \qquad (\star\star)$$

Suppose we are interested in estimators of the form $(\star\star)$ where $\boldsymbol{\beta}$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\|k(\cdot,X_i)-\sum_{j=1}^{n}\beta_j k(\cdot,X_i)\|_{\mathcal{H}}^2+\lambda\|\boldsymbol{\beta}\|_2^2.$$

It can be shown that $\boldsymbol{\beta}=(\mathbf{K}+n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n$ resulting in an estimator,

$$\check{\mu}_\lambda=\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\hat{\mu}_{\mathbb{P}}=\sum_{i=1}^{n}\frac{\gamma_i}{\gamma_i+\lambda}\langle\hat{\mu}_{\mathbb{P}},\phi_i\rangle_{\mathcal{H}}\phi_i,$$

where $\hat{\Sigma}$ is the empirical covariance operator on $\mathcal{H}$.

Spectral Shrinkage

Theorem

Suppose k is a continuous, bounded kernel on a separable topological space $\mathcal{X}$. If $1 \in \mathcal{H}$, then

$$\|\check{\mu}_\lambda - \mu_{\mathbb{P}}\|_{\mathcal{H}} = O_{\mathbb{P}}(n^{-1/2})$$

as $n \to \infty$.

- λ can be estimated using LOOCV.
- Oracle inequality: Open question.

Experiments

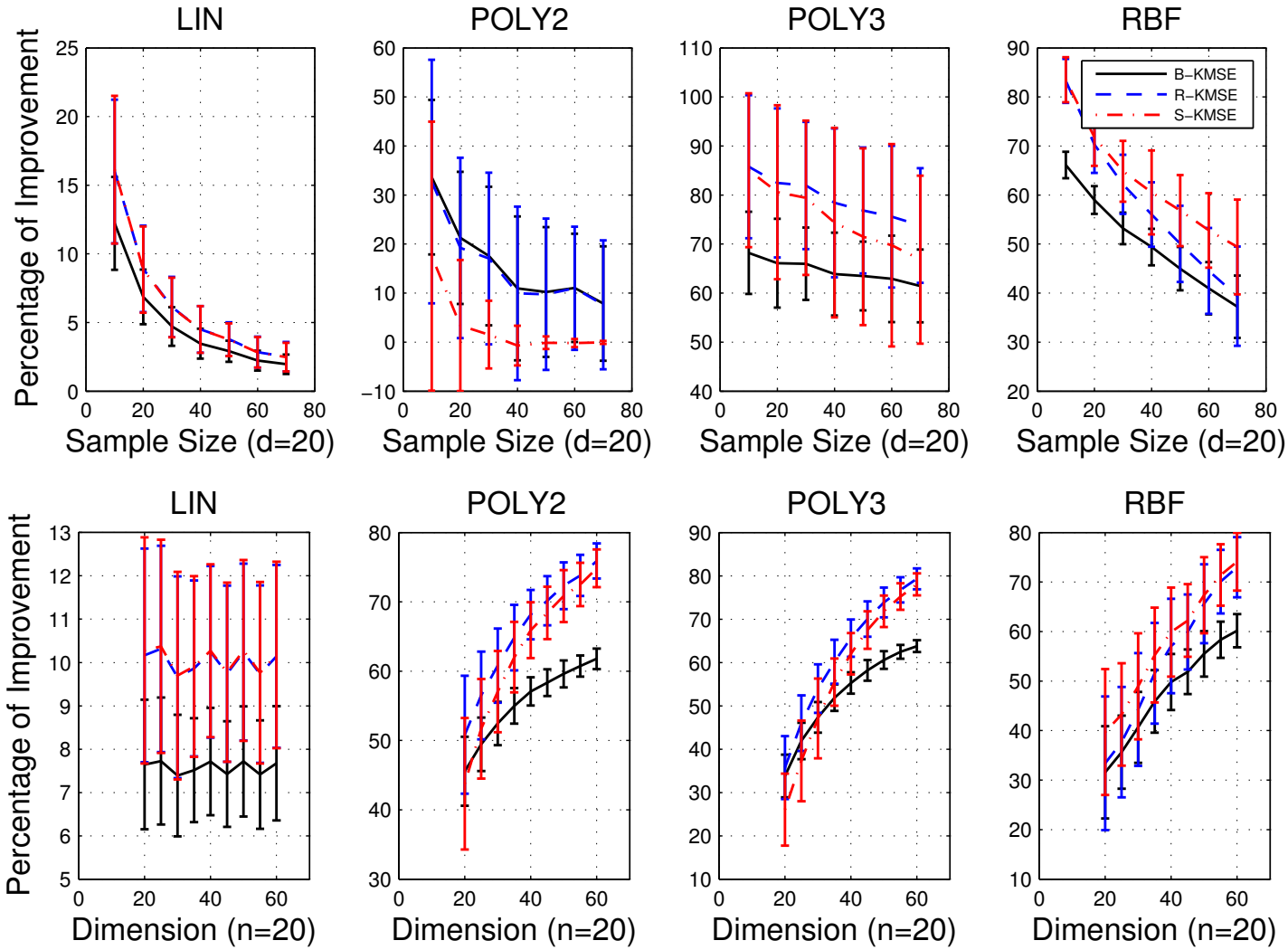


Figure : The percentage of improvement for $\check{\mu}_{\hat{\alpha}}$ (B-KMSE) and $\check{\mu}_{\hat{\lambda}}$ (R-KMSE) over $\hat{\mu}_{\mathbb{P}}$ with varying sample size (n) and dimension (d) over 30 different distributions.

Parzen Window Classifier

Given $\{(x_1, +1), \ldots, (x_m, +1), (y_1, -1), \ldots, (y_m, -1)\}$, the Parzen window classifier is given by

$$f(z) = \operatorname{sgn}\left(\|k(\cdot, z) - \hat{\mu}_{-1}\|_{\mathcal{H}}^2 - \|k(\cdot, z) - \hat{\mu}_{+1}\|_{\mathcal{H}}^2\right)$$

Dataset	KME	B-KMSE	R-KMSE	S-KMSE
Climate Model	0.0348±0.0118	0.0348±0.0118	0.0348±0.0118	0.0348±0.0118
Ionosphere	0.2873±0.0343	**0.2768±0.0359**	**0.2749±0.0341**	0.2800±0.0367
Parkinsons	0.1318±0.0441	0.1250±0.0366	**0.1157±0.0395**	0.1309±0.0396
Pima	0.2951±0.0462	0.2921±0.0442	0.2937±0.0458	0.2943±0.0471
SPECTF	0.2583±0.0829	0.2597±0.0817	**0.2263±0.0626**	0.2417±0.0651
Iris	0.1079±0.0379	0.1071±0.0389	0.1055±0.0389	0.1040±0.0383
Wine	0.1301±0.0381	**0.1183±0.0445**	**0.1161±0.0414**	**0.1183±0.0431**

Table : The classification error rate of Parzen window classifier via different kernel mean estimators. The boldface represents the result whose difference from the baseline, i.e., KME, is statistically significant.

Open Questions

- Alternate ways to construct shrinkage estimators
  - $\breve{\mu} = (1 - g(\|\hat{\mu}_{\mathbb{P}}\|_{\mathcal{H}}))\,\hat{\mu}_{\mathbb{P}}$
  - Choice of g
  - Oracle inequality
  - Reduction to Stein's result
- Minimaxity and admissibility
- Shrinkage estimators for covariance operator

Summary

- Shrinkage estimators for the kernel mean
- Matches the behavior of James-Stein estimator for $d \geq 3$
- Relation of shrinkage to regularization
- Shrinkage using spectral information

Thank You

References I

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

Fukumizu, K., Song, L., and Gretton, A. (2013).
Kernel Bayes' rule: Bayesian inference with positive definite kernels.
Journal of Machine Learning Research, 14:3753–3783.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. (2007).
A kernel method for the two sample problem.
In *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008).
A kernel statistical test of independence.
In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press.

James, W. and Stein, J. (1961).
Estimation with quadratic loss.
In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379. University of California Press.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015).
Towards a learning theory of cause-effect inference.
In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*.

Smola, A. J., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Proc. 18th International Conference on Algorithmic Learning Theory*, pages 13–31. Springer-Verlag, Berlin, Germany.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012).
Feature selection via dependence maximization.
Journal of Machine Learning Research, 13:1393–1434.

Sriperumbudur, B. K. (2011).
Mixture density estimation via Hilbert space embedding of measures.
In *Proceedings of International Symposium on Information Theory*, pages 1027–1030.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. K. (2015).
Two-stage sampled learning theory on distributions.
In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38, pages 948–957. JMLR Workshop and Conference Proceedings.

References II

Tolstikhin, I., Sriperumbudur, B. K., and Muandet, K. (2016).
Minimax estimation of kernel mean embeddings.
arXiv:1602.04361.

Wendland, H. (2005).
Scattered Data Approximation.
Cambridge University Press, Cambridge, UK.