Journal of Machine Learning Research 17 (2016) 1-41 Submitted 5/14; Revised 9/15; Published 4/16

Kernel Mean Shrinkage Estimators

Krikamol Muandet* KRIKAMOL@TUEBINGEN.MPG.DE
Empirical Inference Department, Max Planck Institute for Intelligent Systems
Spemannstraße 38, Tübingen 72076, Germany

Bharath Sriperumbudur* BKS18@PSU.EDU
Department of Statistics, Pennsylvania State University
University Park, PA 16802, USA

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics
10-3 Midoricho, Tachikawa, Tokyo 190-8562 Japan

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit, CSML, University College London
Alexandra House, 17 Queen Square, London - WC1N 3AR, United Kingdom
ORCID 0000-0003-3169-7624

Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Empirical Inference Department, Max Planck Institute for Intelligent Systems
Spemannstraße 38, Tübingen 72076, Germany

Editor: Ingo Steinwart

Abstract

A mean function in a reproducing kernel Hilbert space (RKHS), or a kernel mean, is central to kernel methods in that it is used by many classical algorithms such as kernel principal component analysis, and it also forms the core inference step of modern kernel methods that rely on embedding probability distributions in RKHSs. Given a finite sample, an empirical average has been used commonly as a standard estimator of the true kernel mean. Despite a widespread use of this estimator, we show that it can be improved thanks to the well-known Stein phenomenon. We propose a new family of estimators called kernel mean shrinkage estimators (KMSEs), which benefit from both theoretical justifications and good empirical performance. The results demonstrate that the proposed estimators outperform the standard one, especially in a "large d, small n" paradigm.

Keywords: covariance operator, James-Stein estimators, kernel methods, kernel mean, shrinkage estimators, Stein effect, Tikhonov regularization

1. Introduction

This paper aims to improve the estimation of the mean function in a reproducing kernel Hilbert space (RKHS), or a kernel mean, from a finite sample. A kernel mean is defined with respect to a probability distribution $\mathbb{P}$ over a measurable space $\mathcal{X}$ by

$$\mu_{\mathbb{P}} \triangleq \int_{\mathcal{X}} k(x, \cdot)\ \mathrm{d}\mathbb{P}(x) \in \mathcal{H}, \tag{1}$$

∗. Contributed equally

where $\mu_{\mathbb{P}}$ is a Bochner integral (see, *e.g.*, Diestel and Uhl (1977, Chapter 2) and Dinculeanu (2000, Chapter 1) for a definition of Bochner integral) and $\mathcal{H}$ is a separable RKHS endowed with a measurable reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\int_{\mathcal{X}} \sqrt{k(x,x)}\, \mathrm{d}\mathbb{P}(x) < \infty$.[1] Given an i.i.d sample $x_1, x_2, \ldots, x_n$ from $\mathbb{P}$, the most natural estimate of the true kernel mean is empirical average

$$\hat{\mu}_{\mathbb{P}} \triangleq \frac{1}{n} \sum_{i=1}^{n} k(x_i, \cdot)\,. \tag{2}$$

We refer to this estimator as a *kernel mean estimator* (KME). Though it is the most commonly used estimator of the true kernel mean, the key contribution of this work is to show that there exist estimators that can improve upon this standard estimator.

The kernel mean has recently gained attention in the machine learning community, thanks to the introduction of Hilbert space embedding for distributions (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007). Representing the distribution as a mean function in the RKHS has several advantages. First, if the kernel k is *characteristic*, the map $\mathbb{P} \mapsto \mu_{\mathbb{P}}$ is injective.[2] That is, it preserves all information about the distribution (Fukumizu et al., 2004; Sriperumbudur et al., 2008). Second, basic operations on the distribution can be carried out by means of inner products in RKHS, *e.g.*, $\mathbb{E}_{\mathbb{P}}[f(x)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, which is an essential step in probabilistic inference (see, e.g., Song et al., 2011). Lastly, no intermediate density estimation is required, for example, when testing for homogeneity from finite samples. Thus, the algorithms become less susceptible to the curse of dimensionality; see, *e.g.*, Wasserman (2006, Section 6.5) and Sriperumbudur et al. (2012).

The aforementioned properties make Hilbert space embedding of distributions appealing to many algorithms in modern kernel methods, namely, two-sample testing via maximum mean discrepancy (MMD) (Gretton et al., 2007, 2012), kernel independence tests (Gretton et al., 2008), Hilbert space embedding of HMMs (Song et al., 2010), and kernel Bayes rule (Fukumizu et al., 2011). The performance of these algorithms relies directly on the quality of the empirical estimate $\hat{\mu}_{\mathbb{P}}$.

In addition, the kernel mean has played much more fundamental role as a basic building block of many kernel-based learning algorithms (Vapnik, 1998; Schölkopf et al., 1998). For instance, nonlinear component analyses, such as kernel principal component analysis (KPCA), kernel Fisher discriminant analysis (KFDA), and kernel canonical correlation analysis (KCCA), rely heavily on mean functions and covariance operators in RKHS (Schölkopf et al., 1998; Fukumizu et al., 2007). The kernel K-means algorithm performs clustering in feature space using mean functions as representatives of the clusters (Dhillon et al., 2004). Moreover, the kernel mean also served as a basis in early development of algorithms for classification, density estimation, and anomaly detection (Shawe-Taylor and Cristianini, 2004, Chapter 5). All of these employ the empirical average in (2) as an estimate of the true kernel mean.

1. The separability of $\mathcal{H}$ and measurability of k ensures that $k(\cdot, x)$ is a $\mathcal{H}$-valued measurable function for all $x \in \mathcal{X}$ (Steinwart and Christmann, 2008, Lemma A.5.18). The separability of $\mathcal{H}$ is guaranteed by choosing $\mathcal{X}$ to be a separable topological space and k to be continuous (Steinwart and Christmann, 2008, Lemma 4.33).
2. The notion of characteristic kernel is closely related to the notion of universal kernel. In brief, if the kernel is universal, it is also characteristic, but the reverse direction is not necessarily the case. See, *e.g.*, Sriperumbudur et al. (2011), for more detailed accounts on this topic.

We show in this work that the empirical estimator in (2) is, in a certain sense, not optimal, i.e., there exist "better" estimators (more below), and then propose simple estimators that outperform the empirical estimator. While it is reasonable to argue that $\hat{\mu}_{\mathbb{P}}$ is the "best" possible estimator of $\mu_{\mathbb{P}}$ if nothing is known about $\mathbb{P}$ (in fact $\hat{\mu}_{\mathbb{P}}$ is minimax in the sense of van der Vaart (1998, Theorem 25.21, Example 25.24)), in this paper we show that "better" estimators of $\mu_{\mathbb{P}}$ can be constructed if mild assumptions are made on $\mathbb{P}$. This work is to some extent inspired by Stein's seminal work in 1955, which showed that the maximum likelihood estimator (MLE) of the mean, θ of a multivariate Gaussian distribution $\mathcal{N}(\theta, \sigma^2\mathbf{I})$ is "inadmissible" (Stein, 1955)—i.e., there exists a better estimator—though it is minimax optimal. In particular, Stein showed that there exists an estimator that always achieves smaller total mean squared error regardless of the true $\theta \in \mathbb{R}^d$, when $d \geq 3$. Perhaps the best known estimator of such kind is James-Steins estimator (James and Stein, 1961). Formally, if $X \sim \mathcal{N}(\theta, \sigma^2\mathbf{I})$ with $d \geq 3$, the estimator $\delta(X) = X$ for θ is inadmissible in mean squared sense and is dominated by the following estimator

$$\delta_{\mathrm{JS}}(X) = \left(1 - \frac{(d-2)\sigma^2}{\|X\|^2}\right) X, \tag{3}$$

i.e., $\mathbb{E}\|\delta_{\mathrm{JS}}(X) - \theta\|^2 \leq \mathbb{E}\|\delta(X) - \theta\|^2$ for all θ and there exists at least one θ for which $\mathbb{E}\|\delta_{\mathrm{JS}}(X) - \theta\|^2 < \mathbb{E}\|\delta(X) - \theta\|^2$.

Interestingly, the James-Stein estimator is itself inadmissible, and there exists a wide class of estimators that outperform the MLE, see, *e.g.*, Berger (1976). Ultimately, Stein's result suggests that one can construct estimators better than the usual empirical estimator if the relevant parameters are estimated jointly and if the definition of risk ultimately looks at all of these parameters (or coordinates) together. This finding is quite remarkable as it is counter-intuitive as to why joint estimation should yield better estimators when all parameters are mutually independent (Efron and Morris, 1977). Although the Stein phenomenon has been extensively studied in the statistics community, it has not received much attention in the machine learning community.

The James-Stein estimator is a special case of a larger class of estimators known as *shrinkage estimators* (Gruber, 1998). In its most general form, the shrinkage estimator is a combination of a model with low bias and high variance, and a model with high bias but low variance. For example, one might consider the following estimator:

$$\hat{\theta}_{\mathrm{shrink}} \triangleq \lambda\tilde{\theta} + (1-\lambda)\hat{\theta}_{\mathrm{ML}},$$

where $\lambda \in [0, 1]$, $\hat{\theta}_{\mathrm{ML}}$ denotes the usual maximum likelihood estimate of θ, and $\tilde{\theta}$ is an arbitrary point in the input space. In the case of James-Stein estimator, we have $\tilde{\theta} = 0$. Our proposal of shrinkage estimator to estimate $\mu_{\mathbb{P}}$ will rely on the same principle, but will differ fundamentally from the Stein's seminal works and those along this line in two aspects. First, our setting is "non-parametric" in the sense that we do not assume any parametric form for the distribution, whereas most of traditional works focus on some specific distributions, *e.g.*, the Gaussian distribution. The non-parametric setting is very important in most applications of kernel means because it allows us to perform statistical inference without making any assumption on the parametric form of the true distribution $\mathbb{P}$. Second, our setting involves a "non-linear feature map" into a high-dimensional space.

For example, if we use the Gaussian RBF kernel (see (6)), the mean function $\mu_{\mathbb{P}}$ lives in an infinite-dimensional space. As a result, higher moments of the distribution come into play and therefore one cannot adopt Stein's setting straightforwardly as it involves only the first moment. A direct generalization of James-Stein estimator to infinite-dimensional Hilbert space has been considered, for example, in Berger and Wolpert (1983); Mandelbaum and Shepp (1987); Privault and Rveillac (2008). In those works, the parameter to be estimated is assumed to be the mean of a Gaussian measure on the Hilbert space from which samples are drawn. In contrast, our setting involves samples that are drawn from $\mathbb{P}$ defined on an arbitrary measurable space, and not from a Gaussian measure defined on a Hilbert space.

1.1 Contributions

In the following, we present the main contributions of this work.

1. In Section 2.2, we propose kernel mean shrinkage estimators and show that these estimators can theoretically improve upon the standard empirical estimator, $\hat{\mu}_{\mathbb{P}}$ in terms of the mean squared error (see Theorem 1 and Proposition 4), however, requiring the knowledge of the true kernel mean. We relax this condition in Section 2.3 (see Theorem 5) where without requiring the knowledge of the true kernel mean, we construct shrinkage estimators that are *uniformly* better (in mean squared error) than the empirical estimator over a class of distributions $\mathcal{P}$. For bounded continuous translation invariant kernels, we show that $\mathcal{P}$ reduces to a class of distributions whose characteristic functions have an L^2-norm bounded by a given constant. Through concrete choices for $\mathcal{P}$ in Examples 1 and 2, we discuss the implications of the proposed estimator.

2. While the proposed estimators in Section 2.2 and 2.3 are theoretically interesting, they are not useful in practice as they require the knowledge of the true data generating distribution. In Section 2.4 (see Theorem 7), we present a completely data-dependent estimator (say $\check{\mu}_{\mathbb{P}}$)—referred to as B-KMSE—that is $\sqrt{n}$-consistent and satisfies

$$\mathbb{E}\|\check{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 < \mathbb{E}\|\hat{\mu}_{\mathbb{P}} - \mu_{\mathbb{P}}\|_{\mathcal{H}}^2 + O(n^{-3/2}) \text{ as } n \to \infty. \tag{4}$$

3. In Section 3, we present a regularization interpretation for the proposed shrinkage estimator, wherein the shrinkage parameter is shown to be directly related to the regularization parameter. Based on this relation, we present an alternative approach to choosing the shrinkage parameter (different from the one proposed in Section 2.4) through leave-one-out cross-validation, and show that the corresponding shrinkage estimator (we refer to it as R-KMSE) is also $\sqrt{n}$-consistent and satisfies (4).

4. The regularization perspective also sheds light on constructing new shrinkage estimators that incorporate specific information about the RKHS, based on which we present a new $\sqrt{n}$-consistent shrinkage estimator—referred to as S-KMSE—in Section 4 (see Theorem 13 and Remark 14) that takes into account spectral information of the covariance operator in RKHS. We establish the relation of S-KMSE to the problem of learning smooth operators (Grünewälder et al., 2013) on $\mathcal{H}$, and propose a leave-one-out cross-validation method to obtain a data-dependent shrinkage parameter. However, unlike B-KMSE and R-KMSE, it remains an open question as

to whether S-KMSE with a data-dependent shrinkage parameter is consistent and satisfies an inequality similar to (4). The difficulty in answering these questions lies with the complex form of the estimator, $\tilde{\mu}_{\mathbb{P}}$ which is constructed so as to capture the spectral information of the covariance operator.

5. In Section 6, we empirically evaluate the proposed shrinkage estimators of kernel mean on both synthetic data and several real-world scenarios including Parzen window classification, density estimation and discriminative learning on distributions. The experimental results demonstrate the benefits of our shrinkage estimators over the standard one.

While a shorter version of this work already appeared in Muandet et al. (2014a,b)—particularly, the ideas in Sections 2.2, 3 and 4—, this extended version provides a rigorous theoretical treatment (through Theorems 5, 7, 10, 13 and Proposition 15 which are new) for the proposed estimators and also contains additional experimental results.

2. Kernel Mean Shrinkage Estimators

In this section, we first provide some definitions and notation that are used throughout the paper, following which we present a shrinkage estimator of $\mu_{\mathbb{P}}$. The rest of the section presents various properties including the inadmissibility of the empirical estimator.

2.1 Definitions & Notation

For $a \triangleq (a_1, \ldots, a_d) \in \mathbb{R}^d$, $\|a\|_2 \triangleq \sqrt{\sum_{i=1}^d a_i^2}$. For a topological space $\mathcal{X}$, $C(\mathcal{X})$ (*resp.* $C_b(\mathcal{X})$) denotes the space of all continuous (*resp.* bounded continuous) functions on $\mathcal{X}$. For a locally compact Hausdorff space $\mathcal{X}$, $f \in C(\mathcal{X})$ is said to *vanish at infinity* if for every $\epsilon > 0$ the set $\{x : |f(x)| \geq \epsilon\}$ is compact. The class of all continuous f on $\mathcal{X}$ which vanish at infinity is denoted as $C_0(\mathcal{X})$. $M_b(\mathcal{X})$ (*resp.* $M_+^1(\mathcal{X})$) denotes the set of all finite Borel (*resp.* probability) measures defined on $\mathcal{X}$. For $\mathcal{X} \subset \mathbb{R}^d$, $L^r(\mathcal{X})$ denotes the Banach space of r-power ($r \geq 1$) Lebesgue integrable functions. For $f \in L^r(\mathcal{X})$, $\|f\|_{L^r} \triangleq \left(\int_{\mathcal{X}} |f(x)|^r \, \mathrm{d}x\right)^{1/r}$ denotes the L^r-norm of f for $1 \leq r < \infty$. The Fourier transform of $f \in L^1(\mathbb{R}^d)$ is defined as $f^\wedge(\omega) \triangleq (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-\sqrt{-1}\omega^\top x} \, \mathrm{d}x$, $\omega \in \mathbb{R}^d$. The characteristic function of $\mathbb{P} \in M_+^1(\mathbb{R}^d)$ is defined as $\phi_{\mathbb{P}}(\omega) \triangleq \int e^{\sqrt{-1}\omega^\top x} \, \mathrm{d}\mathbb{P}(x)$, $\omega \in \mathbb{R}^d$.

An RKHS over a set $\mathcal{X}$ is a Hilbert space $\mathcal{H}$ consisting of functions on $\mathcal{X}$ such that for each $x \in \mathcal{X}$ there is a function $k_x \in \mathcal{H}$ with the property

$$\langle f, k_x \rangle_{\mathcal{H}} = f(x), \quad \forall f \in \mathcal{H}. \tag{5}$$

The function $k_x(\cdot) \triangleq k(x, \cdot)$ is called the *reproducing kernel* of $\mathcal{H}$ and the equality (5) is called the *reproducing property* of $\mathcal{H}$. The space $\mathcal{H}$ is endowed with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. Any symmetric and positive semi-definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ uniquely determines an RKHS (Aronszajn, 1950). One of the most popular kernel functions is the Gaussian radial basis function (RBF) kernel on $\mathcal{X} = \mathbb{R}^d$,

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \quad x, y \in \mathcal{X}, \tag{6}$$

where $\|\cdot\|_2$ denotes the Euclidean norm and $\sigma > 0$ is the bandwidth. For $x \in \mathcal{H}_1$ and $y \in \mathcal{H}_2$, $x \otimes y$ denotes the tensor product of x and y, and can be seen as an operator from $\mathcal{H}_2$ to $\mathcal{H}_1$ as $(x \otimes y)z = x\langle y, z\rangle_{\mathcal{H}_2}$ for any $z \in \mathcal{H}_2$, where $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces.

We assume throughout the paper that we observe a sample $x_1, x_2, \ldots, x_n \in \mathcal{X}$ of size n drawn independently and identically (i.i.d.) from some unknown distribution $\mathbb{P}$ defined over a separable topological space $\mathcal{X}$. Denote by μ and $\hat{\mu}$ the true kernel mean (1) and its empirical estimate (2) respectively. We remove the subscript for ease of notation, but we will use $\mu_{\mathbb{P}}$ (resp. $\hat{\mu}_{\mathbb{P}}$) and μ (resp. $\hat{\mu}$) interchangeably. For the well-definedness of μ as a Bochner integral, throughout the paper we assume that k is continuous and $\int_{\mathcal{X}} k(x,x)\, \mathrm{d}\mathbb{P}(x) < \infty$ (see Footnote 1). We measure the quality of an estimator $\tilde{\mu} \in \mathcal{H}$ of μ by the risk function, $R : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, $R(\mu, \tilde{\mu}) = \mathbb{E}\|\mu - \tilde{\mu}\|^2_{\mathcal{H}}$, where $\mathbb{E}$ denotes the expectation over the choice of random sample of size n drawn i.i.d. from the distribution $\mathbb{P}$. When $\tilde{\mu} = \hat{\mu}$, for the ease of notation, we will use Δ to denote $R(\mu, \hat{\mu})$, which can be rewritten as

$$\begin{aligned}
\Delta = \mathbb{E}\|\hat{\mu} - \mu\|^2_{\mathcal{H}} &= \mathbb{E}\|\hat{\mu}\|^2_{\mathcal{H}} - \|\mu\|^2_{\mathcal{H}} = \frac{1}{n^2}\sum_{i,j=1}^{n} \mathbb{E}_{x_i,x_j} k(x_i, x_j) - \|\mu\|^2_{\mathcal{H}} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}_{x_i} k(x_i, x_i) + \frac{1}{n^2}\sum_{i \neq j}^{n} \mathbb{E}_{x_i,x_j} k(x_i, x_j) - \|\mu\|^2_{\mathcal{H}} \\
&= \frac{1}{n}\left(\mathbb{E}_x k(x,x) - \mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})\right), \qquad (7)
\end{aligned}$$

where $\|\mu\|^2_{\mathcal{H}} = \mathbb{E}_{x,\tilde{x}}[k(x,\tilde{x})] \triangleq \mathbb{E}_{x\sim\mathbb{P}}[\mathbb{E}_{\tilde{x}\sim\mathbb{P}}[k(x,\tilde{x})]]$ with x and $\tilde{x}$ being independent copies. An estimator $\hat{\mu}_1$ is said to be *as good as* $\hat{\mu}_2$ if $R(\mu, \hat{\mu}_1) \leq R(\mu, \hat{\mu}_2)$ for any $\mathbb{P}$, and is *better than* $\hat{\mu}_2$ if it is as good as $\hat{\mu}_2$ and $R(\mu, \hat{\mu}_1) < R(\mu, \hat{\mu}_2)$ for at least one $\mathbb{P}$. An estimator is said to be *inadmissible* if there exists a better estimator.

2.2 Shrinkage Estimation of $\mu_{\mathbb{P}}$

We propose the following kernel mean estimator

$$\hat{\mu}_\alpha \triangleq \alpha f^* + (1-\alpha)\hat{\mu} \qquad (8)$$

where $\alpha \geq 0$ and f^* is a fixed, but arbitrary function in $\mathcal{H}$. Basically, it is a shrinkage estimator that shrinks the empirical estimator toward a function f^* by an amount specified by α. The choice of f^* can be arbitrary, but we will assume that f^* is chosen independent of the sample. If $\alpha = 0$, the estimator $\hat{\mu}_\alpha$ reduces to the empirical estimator $\hat{\mu}$. We denote by Δ_α the risk of the shrinkage estimator in (8), *i.e.*, $\Delta_\alpha \triangleq R(\mu, \hat{\mu}_\alpha)$.

Our first theorem asserts that the shrinkage estimator $\hat{\mu}_\alpha$ achieves smaller risk than that of the empirical estimator $\hat{\mu}$ given an appropriate choice of α, regardless of the function f^*.

Theorem 1 *Let $\mathcal{X}$ be a separable topological space. Then for all distributions $\mathbb{P}$ and continuous kernel k satisfying $\int k(x,x)\, \mathrm{d}\mathbb{P}(x) < \infty$, $\Delta_\alpha < \Delta$ if and only if*

$$\alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \mu\|^2_{\mathcal{H}}}\right). \qquad (9)$$

In particular, $\arg\min_{\alpha\in\mathbb{R}}(\Delta_\alpha - \Delta)$ *is unique and is given by* $\alpha_* \triangleq \frac{\Delta}{\Delta + \|f^* - \mu\|^2_{\mathcal{H}}}$.

Proof Note that

$$\Delta_\alpha = \mathbb{E}\|\hat{\mu}_\alpha - \mu\|_{\mathcal{H}}^2 = \|\mathbb{E}[\hat{\mu}_\alpha] - \mu\|_{\mathcal{H}}^2 + \mathbb{E}\,\|\hat{\mu}_\alpha - \mathbb{E}\hat{\mu}_\alpha\|_{\mathcal{H}}^2 = \|\mathrm{Bias}(\hat{\mu}_\alpha)\|_{\mathcal{H}}^2 + \mathrm{Var}(\hat{\mu}_\alpha),$$

where

$$\mathrm{Bias}(\hat{\mu}_\alpha) = \mathbb{E}[\hat{\mu}_\alpha] - \mu = \mathbb{E}[\alpha f^* + (1-\alpha)\hat{\mu}] - \mu = \alpha(f^* - \mu)$$

and

$$\mathrm{Var}(\hat{\mu}_\alpha) = (1-\alpha)^2\mathbb{E}\,\|\hat{\mu} - \mu\|_{\mathcal{H}}^2 = (1-\alpha)^2\Delta.$$

Therefore,

$$\Delta_\alpha = \alpha^2\,\|f^* - \mu\|_{\mathcal{H}}^2 + (1-\alpha)^2\Delta, \tag{10}$$

i.e., $\Delta_\alpha - \Delta = \alpha^2\left[\Delta + \|f^* - \mu\|_{\mathcal{H}}^2\right] - 2\alpha\Delta$. This is clearly negative if and only if (9) holds and is uniquely minimized at $\alpha_* \triangleq \frac{\Delta}{\Delta + \|f^* - \mu\|_{\mathcal{H}}^2}$. ■

Remark 2 *(i) The shrinkage estimator always improves upon the standard one regardless of the direction of shrinkage, as specified by the choice of f^*. In other words, there exists a wide class of kernel mean estimators that achieve smaller risk than the standard one.*

(ii) The range of α depends on the choice of f^*. The further f^* is from μ, the smaller the range of α becomes. Thus, the shrinkage gets smaller if f^* is chosen such that it is far from the true kernel mean. This effect is akin to James-Stein estimator.

(iii) From (9), since $0 < \alpha < 2$, i.e., *$0 < (1-\alpha)^2 < 1$, it follows that $\mathrm{Var}(\hat{\mu}_\alpha) < \mathrm{Var}(\hat{\mu}) = \Delta$,* i.e., *the shrinkage estimator always improves upon the empirical estimator in terms of the variance. Further improvement can be gained by reducing the bias by incorporating the prior knowledge about the location of μ via f^*. This implies that we can potentially gain "twice" by adopting the shrinkage estimator: by reducing variance of the estimator and by incorporating prior knowledge in choosing f^* such that it is close to the true kernel mean.*

While Theorem 1 shows $\hat{\mu}$ to be inadmissible by providing a family of estimators that are better than $\hat{\mu}$, the result is not useful as all these estimators require the knowledge of μ (which is the parameter of interest) through the range of α given in (9). In Section 2.3, we investigate Theorem 1 and show that $\hat{\mu}_\alpha$ can be constructed under some weak assumptions on $\mathbb{P}$, without requiring the knowledge of μ. From (9), the existence of positive α is guaranteed if and only if the risk of the empirical estimator is non-zero. Under some assumptions on k, the following result shows that $\Delta = 0$ if and only if the distribution $\mathbb{P}$ is a Dirac distribution, *i.e.*, the distribution $\mathbb{P}$ is a point mass. This result ensures, in many non-trivial cases, a non-empty range of α for which $\Delta_\alpha - \Delta < 0$.

Proposition 3 *Let $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ be a characteristic kernel where $\psi \in C_b(\mathbb{R}^d)$ is positive definite. Then $\Delta = 0$ if and only if $\mathbb{P} = \delta_x$ for some $x \in \mathbb{R}^d$.*

Proof See Section 5.1. ■

2.2.1 Positive-part Shrinkage Estimator

Similar to James-Stein estimator, we can show that the positive-part version of $\hat{\mu}_\alpha$ also outperforms $\hat{\mu}$, where the positive-part estimator is defined by

$$\hat{\mu}_\alpha^+ \triangleq \alpha f^* + (1-\alpha)_+ \hat{\mu} \tag{11}$$

with $(a)_+ \triangleq a$ if $a > 0$ and zero otherwise. Equation (11) can be rewritten as

$$\hat{\mu}_\alpha^+ = \begin{cases} \alpha f^* + (1-\alpha)\hat{\mu}, & 0 \leq \alpha \leq 1 \\ \alpha f^* & 1 < \alpha < 2. \end{cases} \tag{12}$$

Let $\Delta_\alpha^+ \triangleq \mathbb{E}\|\hat{\mu}_\alpha^+ - \mu\|_{\mathcal{H}}^2$ be the risk of the positive-part estimator. Then, the following result shows that $\Delta_\alpha^+ \leq \Delta_\alpha$, given that α satisfies (9).

Proposition 4 *For any α satisfying (9), we have that $\Delta_\alpha^+ \leq \Delta_\alpha < \Delta$.*

Proof According to (12), we decompose the proof into two parts. First, if $0 \leq \alpha \leq 1$, $\hat{\mu}_\alpha$ and $\hat{\mu}_\alpha^+$ behave exactly the same. Thus, $\Delta_\alpha^+ = \Delta_\alpha$. On the other hand, when $1 < \alpha < 2$, the bias-variance decomposition of these estimators yields

$$\Delta_\alpha = \alpha^2 \|f^* - \mu\|_{\mathcal{H}}^2 + (1-\alpha)^2 \mathbb{E}\|\hat{\mu} - \mu\|_{\mathcal{H}}^2 \quad \text{and} \quad \Delta_\alpha^+ = \alpha^2 \|f^* - \mu\|_{\mathcal{H}}^2.$$

It is easy to see that $\Delta_\alpha^+ < \Delta_\alpha$ when $1 < \alpha < 2$. This concludes the proof. ■

Proposition 4 implies that, when estimating α, it is better to restrict the value of α to be smaller than 1, although it can be greater than 1, as suggested by Theorem 1. The reason is that if $0 \leq \alpha \leq 1$, the bias is an increasing function of α, whereas the variance is a decreasing function of α. On the other hand, if $\alpha > 1$, both bias and variance become increasing functions of α. We will see later in Section 3 that $\hat{\mu}_\alpha$ and $\hat{\mu}_\alpha^+$ can be obtained naturally as a solution to a regularized regression problem.

2.3 Consequences of Theorem 1

As mentioned before, while Theorem 1 is interesting from the perspective of showing that the shrinkage estimator, $\hat{\mu}_\alpha$ performs better—in the mean squared sense—than the empirical estimator, it unfortunately relies on the fact that $\mu_{\mathbb{P}}$ (*i.e.*, the object of interest) is known, which makes $\hat{\mu}_\alpha$ uninteresting. Instead of knowing $\mu_{\mathbb{P}}$, which requires the knowledge of $\mathbb{P}$, in this section, we show that a shrinkage estimator can be constructed that performs better than the empirical estimator, uniformly over a class of probability distributions. To this end, we introduce the notion of an oracle upper bound.

Let $\mathcal{P}$ be a class of probability distributions $\mathbb{P}$ defined on a measurable space $\mathcal{X}$. We define an oracle upper bound as

$$U_{k,\mathcal{P}} \triangleq \inf_{\mathbb{P} \in \mathcal{P}} \frac{2\Delta}{\Delta + \|f^* - \mu\|_{\mathcal{H}}^2}.$$

It follows immediately from Theorem 1 and the definition of $U_{k,\mathcal{P}}$ that if $U_{k,\mathcal{P}} \neq 0$, then for any $\alpha \in (0, U_{k,\mathcal{P}})$, $\Delta_\alpha - \Delta < 0$ holds "uniformly" for all $\mathbb{P} \in \mathcal{P}$. Note that by virtue

of Proposition 3, the class $\mathcal{P}$ cannot contain the Dirac measure δ_x (for any $x \in \mathbb{R}^d$) if the kernel k is translation invariant and characteristic on $\mathbb{R}^d$. Below we give concrete examples of $\mathcal{P}$ for which $U_{k,\mathcal{P}} \neq 0$ so that the above uniformity statement holds. In particular, we show in Theorem 5 below that for $\mathcal{X} = \mathbb{R}^d$, if a non-trivial bound on the L^2-norm of the characteristic function of $\mathbb{P}$ is known, it is possible to construct shrinkage estimators that are better (in mean squared error) than the empirical average. In such a case, unlike in Theorem 1, α does not depend on the individual distribution $\mathbb{P}$, but only on an upper bound associated with a class $\mathcal{P}$.

Theorem 5 *Let $k(x,y) = \psi(x-y)$, $x, y \in \mathbb{R}^d$ with $\psi \in C_b(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$ and ψ is a positive definite function with $\psi(0) > 0$. For a given constant $A \in (0,1)$, let $A_\psi := \frac{A(2\pi)^{d/2}\psi(0)}{\|\psi\|_{L_1}}$ and*

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

where $\phi_{\mathbb{P}}$ denotes the characteristic function of $\mathbb{P}$. Then for all $\mathbb{P} \in \mathcal{P}_{k,A}$, $\Delta_\alpha < \Delta$ if

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

Proof By Theorem 1, we have that

$$\Delta_\alpha < \Delta, \ \forall\ \alpha \in \left(0, \frac{2\Delta}{\Delta + \|f^* - \mu\|_{\mathcal{H}}^2} \right). \tag{13}$$

Consider

$$\begin{aligned}
\frac{\Delta}{\Delta + \|f^* - \mu\|_{\mathcal{H}}^2} &= \frac{\mathbb{E}_x k(x,x) - \mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x) - \mathbb{E}_{x,\tilde{x}} k(x,\tilde{x}) + n\|f^* - \mu\|_{\mathcal{H}}^2} \\
&\overset{(\dagger)}{=} \frac{1 - \frac{\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x)}}{1 + (n-1)\frac{\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x)} + \frac{n\|f^*\|_{\mathcal{H}}^2}{\mathbb{E}_x k(x,x)} - \frac{2n\langle f^*, \mu\rangle_{\mathcal{H}}}{\mathbb{E}_x k(x,x)}} \\
&\geq \frac{1 - \frac{\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x)}}{1 + (n-1)\frac{\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x)} + \frac{n\|f^*\|_{\mathcal{H}}^2}{\mathbb{E}_x k(x,x)} + \frac{2n\|f^*\|_{\mathcal{H}}\sqrt{\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}}{\mathbb{E}_x k(x,x)}},
\end{aligned} \tag{14}$$

where the division by $\mathbb{E}_x k(x,x)$ in $(\dagger)$ is valid since $\mathbb{E}_x k(x,x) = \psi(0) > 0$. Note that the numerator in the r.h.s. of (14) is non-negative since

$$\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x}) \leq \mathbb{E}_x \sqrt{k(x,x)} \mathbb{E}_{\tilde{x}} \sqrt{k(\tilde{x},\tilde{x})} \leq \mathbb{E}_x k(x,x)$$

with equality holding if and only if $\mathbb{P} = \delta_y$ for some $y \in \mathbb{R}^d$ (see Proposition 3). However, for any $A \in (0,1)$ and $y \in \mathbb{R}^d$, it is easy to verify that $\delta_y \notin \mathcal{P}_{k,A}$, which implies the numerator in fact positive. The denominator is clearly positive since $\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x}) \geq 0$ and therefore the r.h.s. of (14) is positive. Also note that

$$\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x}) = \int\int \psi(x-y)\, \mathrm{d}\mathbb{P}(x)\, \mathrm{d}\mathbb{P}(y) \overset{(*)}{=} \int |\phi_{\mathbb{P}}(\omega)|^2 \psi^\wedge(\omega)\, \mathrm{d}\omega$$

$$\leq \sup_{\omega\in\mathbb{R}^d} \psi^\wedge(\omega)\|\phi_{\mathbb{P}}\|_{L_2}^2 \leq (2\pi)^{-d/2}\|\psi\|_{L_1}\|\phi_{\mathbb{P}}\|_{L_2}^2, \tag{15}$$

where $\psi^\wedge$ is the Fourier transform of ψ and $(*)$ follows—see (16) in the proof of Proposition 5 in Sriperumbudur et al. (2011)—by invoking Bochner's theorem (Wendland, 2005, Theorem 6.6), which states that ψ is Fourier transform of a non-negative finite Borel measure with density $(2\pi)^{-d/2}\psi^\wedge$, *i.e.*, $\psi(x) = (2\pi)^{-d/2}\int e^{-ix^\top\omega}\psi^\wedge(\omega)\,\mathrm{d}\omega$, $x\in\mathbb{R}^d$. As $\mathbb{E}_x k(x,x) = \psi(0)$, we have that

$$\frac{\mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})}{\mathbb{E}_x k(x,x)} \leq \frac{A\|\phi_{\mathbb{P}}\|_{L^2}^2}{A_\psi}$$

and therefore for any $\mathbb{P}\in\mathcal{P}_{k,A}$, $\frac{\mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})}{\mathbb{E}_x k(x,x)} \leq A$. Using this in (14) and combining it with (13) yields the result. ■

Remark 6 *(i) Theorem 5 shows that for any $\mathbb{P}\in\mathcal{P}_{k,A}$, it is possible to construct a shrinkage estimator that dominates the empirical estimator,* i.e., *the shrinkage estimator has a strictly smaller risk than that of the empirical estimator.*

(ii) Suppose that $\mathbb{P}$ has a density, denoted by p, with respect to the Lebesgue measure and $\phi_{\mathbb{P}}\in L^2(\mathbb{R}^d)$. By Plancherel's theorem, $p\in L^2(\mathbb{R}^d)$ as $\|p\|_{L_2} = \|\phi_{\mathbb{P}}\|_{L_2}$, which means that $\mathcal{P}_{k,A}$ includes distributions with square integrable densities (note that in general not every p is square integrable). Since

$$\|\phi_{\mathbb{P}}\|_{L_2}^2 = \int |\phi_{\mathbb{P}}(\omega)|^2\,\mathrm{d}\omega \leq \sup_{\omega\in\mathbb{R}^d}|\phi_{\mathbb{P}}(\omega)|\int|\phi_{\mathbb{P}}(\omega)|\,\mathrm{d}\omega = \|\phi_{\mathbb{P}}\|_{L_1},$$

where we used the fact that $\sup_{\omega\in\mathbb{R}^d}|\phi_{\mathbb{P}}(\omega)| = 1$, it is easy to check that

$$\left\{\mathbb{P}\in M_+^1(\mathbb{R}^d) : \|\phi_{\mathbb{P}}\|_{L^1} \leq \frac{A(2\pi)^{d/2}\psi(0)}{\|\psi\|_{L_1}}\right\} \subset \mathcal{P}_{k,A}.$$

This means bounded densities belong to $\mathcal{P}_{k,A}$ as $\phi_{\mathbb{P}}\in L^1(\mathbb{R}^d)$ implies that $\mathbb{P}$ has a density, $p\in C_0(\mathbb{R}^d)$. Moreover, it is easy to check that larger the value of A, larger is the class $\mathcal{P}_{k,A}$ and smaller is the range of α for which $\Delta_\alpha < \Delta$ and vice-versa.

In the following, we present some concrete examples to elucidate Theorem 5.

Example 1 (Gaussian kernel and Gaussian distribution) *Define*

$$\mathcal{N} \triangleq \left\{\mathbb{P}\in M_+^1(\mathbb{R}^d)\,\middle|\,\mathrm{d}\mathbb{P}(x) = \frac{1}{(2\pi\sigma^2)^{d/2}}e^{-\frac{\|x-\theta\|_2^2}{2\sigma^2}}\,\mathrm{d}x,\ \theta\in\mathbb{R}^d,\ \sigma>0\right\},$$

where $\psi(x) = e^{-\|x\|_2^2/2\tau^2}$, $x\in\mathbb{R}^d$ and $\tau>0$. For $\mathbb{P}\in\mathcal{N}$, it is easy to verify that

$$\phi_{\mathbb{P}}(\omega) = e^{\sqrt{-1}\theta^\top\omega - \frac{1}{2}\sigma^2\|\omega\|_2^2},\ \omega\in\mathbb{R}^d \ \text{ and } \ \|\phi_{\mathbb{P}}\|_{L_2}^2 = \int e^{-\sigma^2\|\omega\|_2^2}\,\mathrm{d}\omega = (\pi/\sigma^2)^{d/2}.$$

Also, $\|\psi\|_{L_1} = (2\pi\tau^2)^{d/2}$. Therefore, for $\mathcal{P}_{k,A} \triangleq \{\mathbb{P}\in\mathcal{N} : \sigma^2 \geq \pi\tau^2/A^{2/d}\}$, assuming $f^* = 0$, we obtain the result in Theorem 5, i.e., *the result in Theorem 5 holds for all Gaussian distributions that are smoother (having larger variance) than that of the kernel.*

Example 2 (Linear kernel) *Suppose $f^* = 0$ and $k(x, y) = x^\top y$. While the setting of Theorem 5 does not fit this choice of k, an inspection of its proof shows that it is possible to construct a shrinkage estimator that improves upon $\mu_{\mathbb{P}}$ for an appropriate class of distributions. To this end, let ϑ and Σ represent the mean vector and covariance matrix of a distribution $\mathbb{P}$ defined on $\mathbb{R}^d$. Then it is easy to check that $\frac{\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x)} = \frac{\|\vartheta\|_2^2}{\operatorname{trace}(\Sigma)+\|\vartheta\|_2^2}$ and therefore for a given $A \in (0, 1)$, define*

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

*From (13) and (14), it is clear that for any $\mathbb{P} \in \mathcal{P}_{k,A}$, $\Delta_\alpha < \Delta$ if $\alpha \in \left(0, \frac{2(1-A)}{1+(n-1)A}\right]$. Note that this choice of kernel yields the setting similar to classical James-Stein estimation. In James-Stein estimation, $\mathbb{P} \in \mathcal{N}$ (see Example 1 for the definition of $\mathcal{N}$) and ϑ is estimated as $(1-\tilde{\alpha})\hat{\vartheta}$—which improves upon $\hat{\vartheta}$—where $\tilde{\alpha}$ depends on the sample $(x_i)_{i=1}^n$ and $\hat{\vartheta}$ is the sample mean. In our case, for all $\mathbb{P} \in \mathcal{P}_{k,A} = \left\{ \mathbb{P} \in \mathcal{N} : \|\vartheta\|_2 \leq \sigma\sqrt{\frac{dA}{1-A}} \right\}$, $\Delta_\alpha < \Delta$ if $\alpha \in \left(0, \frac{2(1-A)}{1+(n-1)A}\right]$. In addition, in contrast to the James-stein estimator which improves upon the empirical estimator (*i.e.*, *sample mean) for only $d \geq 3$, we note here that the proposed estimator improves for any d as long as $\mathbb{P} \in \mathcal{P}_{k,A}$. On the other hand, the proposed estimator requires some knowledge about the distribution (particularly a bound on $\|\vartheta\|_2$), which the James-Stein estimator does not (see Section 2.5 for more details).*

2.4 Data-Dependent Shrinkage Parameter

The discussion so far showed that the shrinkage estimator in (8) performs better than the empirical estimator if the data generating distribution satisfies a certain mild condition (see Theorem 5; Examples 1 and 2). However, since this condition is usually not checkable in practice, the shrinkage estimator lacks applicability. In this section, we present a completely data driven shrinkage estimator by estimating the shrinkage parameter α from data so that the estimator does not require any knowledge of the data generating distribution.

Since the maximal difference between Δ_α and Δ occurs at α_* (see Theorem 1), given an i.i.d. sample $X = \{x_1, x_2, \ldots, x_n\}$ from $\mathbb{P}$, we propose to estimate μ using $\hat{\mu}_{\tilde{\alpha}} = (1-\tilde{\alpha})\hat{\mu}$ (*i.e.*, assuming $f^* = 0$) where $\tilde{\alpha}$ is an estimator of $\alpha_* = \Delta/(\Delta + \|\mu\|_{\mathcal{H}}^2)$ given by

$$\tilde{\alpha} = \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\mu}\|_{\mathcal{H}}^2}, \tag{16}$$

with $\hat{\Delta}$ and $\hat{\mu}$ being the empirical versions of Δ and μ, respectively (see Theorem 7 for precise definitions). The following result shows that $\tilde{\alpha}$ is a $n\sqrt{n}$-consistent estimator of α_* and $\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}$ concentrates around $\|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}$. In addition, we show that

$$\Delta_{\alpha_*} \leq \Delta_{\tilde{\alpha}} \leq \Delta_{\alpha_*} + O(n^{-3/2}) \text{ as } n \to \infty,$$

which means the performance of $\hat{\mu}_{\tilde{\alpha}}$ is similar to that of the best estimator (in mean squared sense) of the form $\hat{\mu}_\alpha$. In what follows, we will call the estimator $\hat{\mu}_{\tilde{\alpha}}$ an *empirical-bound kernel mean shrinkage estimator (B-KMSE).*

Theorem 7 *Suppose $n \geq 2$ and $f^* = 0$. Let k be a continuous kernel on a separable topological space $\mathcal{X}$ satisfying $\int_{\mathcal{X}} k(x,x)\,\mathrm{d}\mathbb{P}(x) < \infty$. Define*

$$\hat{\Delta} \triangleq \frac{\hat{\mathbb{E}}k(x,x) - \hat{\mathbb{E}}k(x,\tilde{x})}{n} \quad \text{and} \quad \|\hat{\mu}\|_{\mathcal{H}}^2 \triangleq \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j)$$

where $\hat{\mathbb{E}}k(x,x) \triangleq \frac{1}{n}\sum_{i=1}^{n} k(x_i,x_i)$ and $\hat{\mathbb{E}}k(x,\tilde{x}) \triangleq \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(x_i,x_j)$ are the empirical estimators of $\mathbb{E}_x k(x,x)$ and $\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})$ respectively. Assume there exist finite constants $\kappa_1 > 0$, $\kappa_2 > 0$, $\sigma_1 > 0$ and $\sigma_2 > 0$ such that

$$\mathbb{E}\|k(\cdot,x) - \mu\|_{\mathcal{H}}^m \leq \frac{m!}{2}\sigma_1^2\kappa_1^{m-2}, \quad \forall\, m \geq 2. \tag{17}$$

and

$$\mathbb{E}|k(x,x) - \mathbb{E}_x k(x,x)|^m \leq \frac{m!}{2}\sigma_2^2\kappa_2^{m-2}, \quad \forall\, m \geq 2. \tag{18}$$

Then

$$|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2}) \;\text{ and }\; \left|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right| = O_{\mathbb{P}}(n^{-3/2})$$

as $n \to \infty$. In particular,

$$\min_{\alpha} \mathbb{E}\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 \leq \min_{\alpha} \mathbb{E}\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}) \tag{19}$$

as $n \to \infty$.

Proof See Section 5.2. ■

Remark 8 *(i) $\hat{\mu}_{\tilde{\alpha}}$ is a $\sqrt{n}$-consistent estimator of μ. This follows from*

$$\begin{aligned}\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} &\leq \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}} + O_{\mathbb{P}}(n^{-3/2}) \\ &\leq (1-\alpha_*)\|\hat{\mu} - \mu\|_{\mathcal{H}} + \alpha_*\|\mu\|_{\mathcal{H}} + O_{\mathbb{P}}(n^{-3/2})\end{aligned}$$

with

$$\alpha_* = \frac{\Delta}{\Delta + \|\mu\|_{\mathcal{H}}^2} = \frac{\mathbb{E}_x k(x,x) - \mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})}{\mathbb{E}_x k(x,x) + (n-1)\mathbb{E}_{x,\tilde{x}} k(x,\tilde{x})} = O(n^{-1})$$

as $n \to \infty$. Using (38), we obtain $\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} = O_{\mathbb{P}}(n^{-1/2})$ as $n \to \infty$, which implies that $\hat{\mu}_{\tilde{\alpha}}$ is a $\sqrt{n}$-consistent estimator of μ.

(ii) Equation (19) shows that $\Delta_{\tilde{\alpha}} \leq \Delta_{\alpha_*} + O(n^{-3/2})$ where $\Delta_{\alpha_*} < \Delta$ (see Theorem 1) and therefore for any $\mathbb{P}$ satisfying (17) and (18), $\Delta_{\tilde{\alpha}} < \Delta + O(n^{-3/2})$ as $n \to \infty$.

(iii) Suppose the kernel is bounded, i.e., *$\sup_{x,y\in\mathcal{X}} |k(x,y)| \leq \kappa < \infty$. Then it is easy to verify that (17) and (18) hold with $\sigma_1 = \sqrt{\kappa}$, $\kappa_1 = 2\sqrt{\kappa}$, $\sigma_2 = \kappa$ and $\kappa_2 = 2\kappa$ and therefore the claims in Theorem 7 hold for bounded kernels.*

(iv) For $k(x,y) = x^\top y$, we have

$$\mathbb{E}\|k(\cdot,x) - \mu\|_{\mathcal{H}}^m = \mathbb{E}\left(\|k(\cdot,x) - \mu\|_{\mathcal{H}}^2\right)^{m/2} = \mathbb{E}\left(\|x - \mathbb{E}_x x\|_2^2\right)^{m/2} = \mathbb{E}\|x - \mathbb{E}_x x\|_2^m$$

and

$$\mathbb{E}|k(x,x) - \mathbb{E}_x k(x,x)|^m = \mathbb{E}|\|x\|_2^2 - \mathbb{E}_x\|x\|_2^2|^m.$$

The conditions in (17) and (18) hold for $\mathbb{P} \in \mathcal{N}$ where $\mathcal{N}$ is defined in Example 1. With $\mathbb{P} \in \mathcal{N}$ and $k(x,y) = x^\top y$, the problem of estimating μ reduces to estimating θ, for which we have presented a James-Stein-like estimator, $\hat{\mu}_{\tilde{\alpha}}$ that satisfies the oracle inequality in (19).

(v) While the moment conditions in (17) and (18) are obviously satisfied by bounded kernels, for unbounded kernels, these conditions are quite stringent as they require all the higher moments to exist. These conditions can be weakened and the proof of Theorem 7 can be carried out using Chebyshev inequality instead of Bernstein's inequality but at the cost of a slow rate in (19).

2.5 Connection to James-Stein Estimator

In this section, we explore the connection of our proposed estimator in (8) to the James-Stein estimator. Recall that Stein's setting deals with estimating the mean of the Gaussian distribution $\mathcal{N}(\theta, \sigma^2\mathbf{I}_d)$, which can be viewed as a special case of kernel mean estimation when we restrict to the class of distributions $\mathcal{P} \triangleq \{\mathcal{N}(\theta, \sigma^2\mathbf{I}_d) \,|\, \theta \in \mathbb{R}^d\}$ and a linear kernel $k(x,y) = x^\top y$, $x, y \in \mathbb{R}^d$ (see Example 2). In this case, it is easy to verify that $\Delta = d\sigma^2/n$ and $\Delta_\alpha < \Delta$ for

$$\alpha \in \left(0, \frac{2d\sigma^2}{d\sigma^2 + n\|\theta\|^2}\right).$$

Let us assume that $n = 1$, in which case, we obtain $\Delta_\alpha < \Delta$ for $\alpha \in \left(0, \frac{2d\sigma^2}{\mathbb{E}_x\|x\|^2}\right)$ as $\mathbb{E}_x\|x\|^2 = \|\theta\|^2 + d\sigma^2$. Note that the choice of α is dependent on $\mathbb{P}$ through $\mathbb{E}_x\|x\|^2$ which is not known in practice. To this end, we replace it with the empirical version $\|x\|^2$ that depends only on the sample x. For an arbitrary constant $c \in (0, 2d)$, the shrinkage estimator (assuming $f^* = 0$) can thus be written as

$$\hat{\mu}_\alpha = (1-\alpha)\hat{\mu} = \left(1 - \frac{c\sigma^2}{\|x\|^2}\right) x = x - \frac{c\sigma^2 x}{\|x\|^2},$$

which is exactly the James-Stein estimator in (3). This particular way of estimating the shrinkage parameter α has an intriguing consequence, as shown in Stein's seminal works (Stein, 1955; James and Stein, 1961), that the shrinkage estimator $\hat{\mu}_\alpha$ can be shown to dominate the maximum likelihood estimator $\hat{\mu}$ *uniformly* over all θ.

While it is compelling to see that there is seemingly a fundamental principle underlying both these settings, this connection also reveals crucial difference between our approach and classical setting of Stein—notably, original James-Stein estimator improves upon the sample mean even when α is data-dependent (see $\hat{\mu}_\alpha$ above), however, with the crucial assumption that x is normally distributed.

3. Kernel Mean Estimation as Regression Problem

In Section 2, we have shown that James-Stein-like shrinkage estimator, *i.e.*, Equation (8), improves upon the empirical estimator in estimating the kernel mean. In this section,

we provide a regression perspective to shrinkage estimation. The starting point of the connection between regression and shrinkage estimation is the observation that the kernel mean $\mu_{\mathbb{P}}$ and its empirical estimate $\hat{\mu}_{\mathbb{P}}$ can be obtained as minimizers of the following risk functionals,

$$\mathcal{E}(g) \triangleq \int_{\mathcal{X}} \|k(\cdot, x) - g\|_{\mathcal{H}}^2 \, \mathrm{d}\mathbb{P}(x) \;\text{ and }\; \widehat{\mathcal{E}}(g) \triangleq \frac{1}{n} \sum_{i=1}^{n} \|k(\cdot, x_i) - g\|_{\mathcal{H}}^2,$$

respectively (Kim and Scott, 2012). Given these formulations, it is natural to ask if minimizing the regularized version of $\widehat{\mathcal{E}}(g)$ will give a "better" estimator. While this question is interesting, it has to be noted that in principle, there is really no need to consider a regularized formulation as the problem of minimizing $\widehat{\mathcal{E}}$ is not ill-posed, unlike in function estimation or regression problems. To investigate this question, we consider the minimization of the following regularized empirical risk functional,

$$\widehat{\mathcal{E}}_{\lambda}(g) \triangleq \widehat{\mathcal{E}}(g) + \lambda \Omega(\|g\|_{\mathcal{H}}) = \frac{1}{n} \sum_{i=1}^{n} \|k(\cdot, x_i) - g\|_{\mathcal{H}}^2 + \lambda \Omega(\|g\|_{\mathcal{H}}), \tag{20}$$

where $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ denotes a monotonically increasing function and $\lambda > 0$ is the regularization parameter. By representer theorem (Schölkopf et al., 2001), any function $g \in \mathcal{H}$ that is a minimizer of (20) lies in a subspace spanned by $\{k(\cdot, x_1), \ldots, k(\cdot, x_n)\}$, *i.e.*, $g = \sum_{j=1}^{n} \beta_j k(\cdot, x_j)$ for some $\boldsymbol{\beta} \triangleq [\beta_1, \ldots, \beta_n]^\top \in \mathbb{R}^n$. Hence, by setting $\Omega(\|g\|_{\mathcal{H}}) = \|g\|_{\mathcal{H}}^2$, we can rewrite (20) in terms of $\boldsymbol{\beta}$ as

$$\widehat{\mathcal{E}}(g) + \lambda \Omega(\|g\|_{\mathcal{H}}) = \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} - 2 \boldsymbol{\beta}^\top \mathbf{K} \mathbf{1}_n + \lambda \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} + c, \tag{21}$$

where $\mathbf{K}$ is an $n \times n$ Gram matrix such that $\mathbf{K}_{ij} = k(x_i, x_j)$, c is a constant that does not depend on $\boldsymbol{\beta}$, and $\mathbf{1}_n = [1/n, 1/n, \ldots, 1/n]^\top$. Differentiating (21) with respect to $\boldsymbol{\beta}$ and setting it to zero yields an optimal weight vector $\boldsymbol{\beta} = \left(\frac{1}{1+\lambda}\right) \mathbf{1}_n$ and so the minimizer of (20) is given by

$$\hat{\mu}_{\lambda} = \frac{1}{1+\lambda} \hat{\mu} = \left(1 - \frac{\lambda}{1+\lambda}\right) \hat{\mu} \triangleq (1 - \alpha) \hat{\mu}, \tag{22}$$

which is nothing but the shrinkage estimator in (8) with $\alpha = \frac{\lambda}{1+\lambda}$ and $f^* = 0$. This provides a nice relation between shrinkage estimation and regularized risk minimization, wherein the regularization helps in shrinking the estimator $\hat{\mu}$ towards zero although it is not required from the point of view of ill-posedness. In particular, since $0 < 1 - \alpha < 1$, $\hat{\mu}_{\lambda}$ corresponds to a *positive-part* estimator proposed in Section 2.2.1 when $f^* = 0$.

Note that $\hat{\mu}_{\lambda}$ is a consistent estimator of μ as $\lambda \to 0$ and $n \to \infty$, which follows from

$$\|\hat{\mu}_{\lambda} - \mu\|_{\mathcal{H}} \leq \frac{1}{1+\lambda} \|\hat{\mu} - \mu\|_{\mathcal{H}} + \frac{\lambda}{1+\lambda} \|\mu\|_{\mathcal{H}} \leq O_{\mathbb{P}}(n^{-1/2}) + O(\lambda).$$

In particular $\lambda = \tau n^{-1/2}$ (for some constant $\tau > 0$) yields the slowest possible rate for $\lambda \to 0$ such that the best possible rate of $n^{-1/2}$ is obtained for $\|\hat{\mu}_{\lambda} - \mu\|_{\mathcal{H}} \to 0$ as $n \to \infty$. In addition, following the idea in Theorem 5, it is easy to show that $\mathbb{E}\|\hat{\mu}_{\lambda} - \mu\|_{\mathcal{H}}^2 < \Delta$ if

$\tau \in \left(0, \frac{2\sqrt{n}\Delta}{\|\mu\|_{\mathcal{H}}^2 - \Delta}\right)$. Note that $\hat{\mu}_\lambda$ is not useful in practice as λ is not known *a priori*. However, by choosing

$$\lambda = \frac{\hat{\Delta}}{\|\hat{\mu}\|_{\mathcal{H}}^2},$$

it is easy to verify (see Theorem 7 and Remark 8) that

$$\mathbb{E}\|\hat{\mu}_\lambda - \mu\|_{\mathcal{H}}^2 < \mathbb{E}\|\hat{\mu} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}) \tag{23}$$

as $n \to \infty$. Owing to the connection of $\hat{\mu}_\lambda$ to a regression problem, in the following, we present an alternate data-dependent choice of λ obtained from leave-one-out cross validation (LOOCV) that also satisfies (23), and we refer to the corresponding estimator as *regularized kernel mean shrinkage estimator (R-KMSE)*.

To this end, for a given shrinkage parameter λ, denote by $\hat{\mu}_\lambda^{(-i)}$ as the kernel mean estimated from $\{x_j\}_{j=1}^n \backslash \{x_i\}$. We will measure the quality of $\hat{\mu}_\lambda^{(-i)}$ by how well it approximates $k(\cdot, x_i)$ with the overall quality being quantified by the cross-validation score,

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left\| k(\cdot, x_i) - \hat{\mu}_\lambda^{(-i)} \right\|_{\mathcal{H}}^2. \tag{24}$$

The LOOCV formulation in (24) differs from the one used in regression, wherein instead of measuring the deviation of the prediction made by the function on the omitted observation, we measure the deviation between the feature map of the omitted observation and the function itself. The following result shows that the shrinkage parameter in $\hat{\mu}_\lambda$ (see (22)) can be obtained analytically by minimizing (24) and requires $O(n^2)$ operations to compute.

Proposition 9 *Let $n \geq 2$, $\rho := \frac{1}{n^2}\sum_{i,j=1}^n k(x_i, x_j)$ and $\varrho := \frac{1}{n}\sum_{i=1}^n k(x_i, x_i)$. Assuming $n\rho > \varrho$, the unique minimizer of $LOOCV(\lambda)$ is given by*

$$\lambda_r = \frac{n(\varrho - \rho)}{(n-1)(n\rho - \varrho)}. \tag{25}$$

Proof See Section 5.3. ∎

It is instructive to compare

$$\alpha_r = \frac{\lambda_r}{\lambda_r + 1} = \frac{\varrho - \rho}{(n-2)\rho + \varrho/n} \tag{26}$$

to the one in (16), where the latter can be shown to be $\frac{\varrho - \rho}{\varrho + (n-2)\rho}$, by noting that $\hat{\mathbb{E}}k(x,x) = \varrho$ and $\hat{\mathbb{E}}k(x,\tilde{x}) = \frac{n\rho - \varrho}{n-1}$ (in Theorem 7, we employ the U-statistic estimator of $\mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})$, whereas ρ in Proposition 9 can be seen as a V-statistic counterpart). This means α_r obtained from LOOCV will be relatively larger than the one obtained from (16). Like in Theorem 7, the requirement that $n \geq 2$ in Theorem 9 stems from the fact that at least two data points are needed to evaluate the LOOCV score. Note that $n\rho > \varrho$ if and only if $\hat{\mathbb{E}}k(x,\tilde{x}) > 0$, which is guaranteed if the kernel is positive valued. We refer to $\hat{\mu}_{\lambda_r}$ as R-KMSE, whose $\sqrt{n}$-consistency is established by the following result, which also shows that $\hat{\mu}_{\lambda_r}$ satisfies (23).

Theorem 10 *Let $n \geq 2$, $n\rho > \varrho$ where ρ and ϱ are defined in Proposition 9 and k satisfies the assumptions in Theorem 7. Then $\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} = O_{\mathbb{P}}(n^{-1/2})$,*

$$\min_{\alpha} \mathbb{E}\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 \leq \mathbb{E}\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}}^2 \leq \min_{\alpha} \mathbb{E}\|\hat{\mu}_{\alpha} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}) \tag{27}$$

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

$$\mathbb{E}\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}}^2 < \mathbb{E}\|\hat{\mu} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2}) \tag{28}$$

as $n \to \infty$.

Proof See Section 5.4. ■

4. Spectral Shrinkage Estimators

Consider the following regularized risk minimization problem

$$\arg\inf_{\mathbf{F} \in \mathcal{H}\otimes\mathcal{H}} \quad \mathbb{E}_{x\sim\mathbb{P}} \|k(x,\cdot) - \mathbf{F}[k(x,\cdot)]\|_{\mathcal{H}}^2 + \lambda \|\mathbf{F}\|_{\mathrm{HS}}^2, \tag{29}$$

where the minimization is carried over the space of Hilbert-Schmidt operators, $\mathbf{F}$ on $\mathcal{H}$ with $\|\mathbf{F}\|_{\mathrm{HS}}$ being the Hilbert-Schmidt norm of $\mathbf{F}$. As an interpretation, we are finding a smooth operator $\mathbf{F}$ that maps $k(x,\cdot)$ to itself (see Grünewälder et al. (2013) for more details on this smooth operator framework). It is not difficult to show that the solution to (29) is given by $\mathbf{F} = \Sigma_{XX}(\Sigma_{XX} + \lambda I)^{-1}$ where $\Sigma_{XX} = \int k(\cdot.x) \otimes k(\cdot, x)\, \mathrm{d}\mathbb{P}(x)$ is a covariance operator defined on $\mathcal{H}$ (see, e.g., Grünewälder et al., 2012). Note that Σ_{XX} is a Bochner integral, which is well-defined as a Hilbert-Schmidt operator if $\mathcal{X}$ is a separable topological space and k is a continuous kernel satisfying $\int k(x,x)\, \mathrm{d}\mathbb{P}(x) < \infty$. Consequently, let us define

$$\mu_{\lambda} = \mathbf{F}\mu = \Sigma_{XX}(\Sigma_{XX} + \lambda I)^{-1}\mu,$$

which is an approximation to μ as it can be shown that $\|\mu_{\lambda} - \mu\|_{\mathcal{H}} \to 0$ as $\lambda \to 0$ (see the proof of Theorem 13). Given an i.i.d. sample $x_1, \ldots, x_n$ from $\mathbb{P}$, the empirical counterpart of (29) is given by

$$\arg\min_{\mathbf{F} \in \mathcal{H}\otimes\mathcal{H}} \quad \frac{1}{n}\sum_{i=1}^{n} \|k(x_i,\cdot) - \mathbf{F}[k(x_i,\cdot)]\|_{\mathcal{H}}^2 + \lambda \|\mathbf{F}\|_{\mathrm{HS}}^2 \tag{30}$$

resulting in

$$\breve{\mu}_{\lambda} \triangleq \mathbf{F}\hat{\mu} = \hat{\Sigma}_{XX}(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\mu} \tag{31}$$

where $\hat{\Sigma}_{XX}$ is the empirical covariance operator on $\mathcal{H}$ given by

$$\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot.x_i) \otimes k(\cdot, x_i).$$

Unlike $\hat{\mu}_{\lambda}$ in (22), $\breve{\mu}_{\lambda}$ shrinks $\hat{\mu}$ differently in each coordinate by taking the eigenspectrum of $\hat{\Sigma}_{XX}$ into account (see Proposition 11) and so we refer to it as the *spectral kernel mean shrinkage estimator (S-KMSE)*.

Proposition 11 *Let $\{(\gamma_i, \phi_i)\}_{i=1}^n$ be eigenvalue and eigenfunction pairs of $\hat{\Sigma}_{XX}$. Then*

$$\check{\mu}_\lambda = \sum_{i=1}^{n} \frac{\gamma_i}{\gamma_i + \lambda} \langle \hat{\mu}, \phi_i \rangle_{\mathcal{H}} \phi_i.$$

Proof Since $\hat{\Sigma}_{XX}$ is a finite rank operator, it is compact. Since it is also a self-adjoint operator on $\mathcal{H}$, by Hilbert-Schmidt theorem (Reed and Simon, 1972, Theorems VI.16, VI.17), we have $\hat{\Sigma}_{XX} = \sum_{i=1}^n \gamma_i \langle \phi_i, \cdot \rangle_{\mathcal{H}} \phi_i$. The result follows by using this in (31). ■

As shown in Proposition 11, the effect of S-KMSE is to reduce the contribution of high frequency components of $\hat{\mu}$ (*i.e.*, contribution of $\hat{\mu}$ along the directions corresponding to smaller γ_i) when $\hat{\mu}$ is expanded in terms of the eigenfunctions of the empirical covariance operator, which are nothing but the kernel PCA basis (Rasmussen and Williams, 2006, Section 4.3). This means, similar to R-KMSE, S-KMSE also shrinks $\hat{\mu}$ towards zero, however, the difference being that while R-KMSE shrinks equally in all coordinates, S-KMSE controls the amount of shrinkage by the information contained in each coordinate. In particular, S-KMSE takes into account more information about the kernel by allowing for different amount of shrinkage in each coordinate direction according to the value of γ_i, wherein the shrinkage is small in the coordinates whose γ_i are large. Moreover, Proposition 11 reveals that the effect of shrinkage is akin to *spectral filtering* (Bauer et al., 2007)—which in our case corresponds to Tikhonov regularization—wherein S-KMSE filters out the high-frequency components of the spectral representation of the kernel mean. Muandet et al. (2014b) leverages this observation and generalizes S-KMSE to a family of shrinkage estimators via spectral filtering algorithms.

The following result presents an alternate representation for $\check{\mu}_\lambda$, using which we relate the smooth operator formulation in (30) to the regularization formulation in (20).

Proposition 12 *Let $\Phi : \mathbb{R}^n \to \mathcal{H}$, $\mathbf{a} \mapsto \sum_{i=1}^n a_i k(\cdot, x_i)$ where $\mathbf{a} \triangleq (a_1, \ldots, a_n)$. Then*

$$\check{\mu}_\lambda = \hat{\Sigma}_{XX}(\hat{\Sigma}_{XX} + \lambda I)^{-1} \hat{\mu} = \Phi(\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n,$$

where $\mathbf{K}$ is the Gram matrix, I is an identity operator on $\mathcal{H}$, $\mathbf{I}$ is an $n \times n$ identity matrix and $\mathbf{1}_n \triangleq [1/n, \ldots, 1/n]^\top$.

Proof See Section 5.5. ■

From Proposition 12, it is clear that

$$\check{\mu}_\lambda = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\boldsymbol{\beta}_s)_j k(\cdot, x_j) \tag{32}$$

where $\boldsymbol{\beta}_s \triangleq \sqrt{n}(\mathbf{K} + n\lambda \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n$. Given the form of $\check{\mu}_\lambda$ in (32), it is easy to verify that $\boldsymbol{\beta}_s$ is the minimizer of (20) when $\widehat{\mathcal{E}}_\lambda$ is minimized over $\{g = \frac{1}{\sqrt{n}} \sum_{j=1}^n (\boldsymbol{\beta})_j k(\cdot, x_j) : \boldsymbol{\beta} \in \mathbb{R}^n\}$ with $\Omega(\|g\|_{\mathcal{H}}) \triangleq \|\boldsymbol{\beta}\|_2^2$.

The following result, discussed in Remark 14, establishes the consistency and convergence rate of S-KMSE, $\check{\mu}_\lambda$.

Theorem 13 *Suppose $\mathcal{X}$ is a separable topological space and k is a continuous, bounded kernel on $\mathcal{X}$. Then the following hold.*

(i) If $\mu \in \overline{\mathcal{R}(\Sigma_{XX})}$, then $\|\check{\mu}_\lambda - \mu\|_{\mathcal{H}} \to 0$ as $\lambda\sqrt{n} \to \infty$, $\lambda \to 0$ and $n \to \infty$.

(ii) If $\mu \in \mathcal{R}(\Sigma_{XX})$, then $\|\check{\mu}_\lambda - \mu\|_{\mathcal{H}} = O_{\mathbb{P}}(n^{-1/2})$ for $\lambda = cn^{-1/2}$ with $c > 0$ being a constant independent of n.

Proof See Section 5.6. ■

Remark 14 *While Theorem 13(i) shows that S-KMSE, $\check{\mu}_\lambda$ is not universally consistent,* i.e., *S-KMSE is not consistent for all $\mathbb{P}$ but only for those $\mathbb{P}$ that satisfies $\mu \in \overline{\mathcal{R}(\Sigma_{XX})}$, under some additional conditions on the kernel, the universal consistency of S-KMSE can be guaranteed. This is achieved by assuming that constant functions are included in $\mathcal{H}$,* i.e., *$1 \in \mathcal{H}$. Note that if $1 \in \mathcal{H}$, then it is easy to check that there exists $g \in \mathcal{H}$ (choose $g = 1$) such that $\mu = \Sigma_{XX} g = \int k(\cdot, x) g(x)\, \mathrm{d}\mathbb{P}(x)$,* i.e., *$\mu \in \mathcal{R}(\Sigma_{XX})$, and, therefore, by Theorem 13, $\check{\mu}_\lambda$ is not only universally consistent but also achieves a rate of $n^{-1/2}$. Choosing $k(x, y) = \tilde{k}(x, y) + b$, $x, y \in \mathcal{X}$, $b > 0$ where $\tilde{k}$ is any bounded, continuous positive definite kernel ensures that $1 \in \mathcal{H}$.*

Note that the estimator $\check{\mu}_\lambda$ requires the knowledge of the shrinkage or regularization parameter, λ. Similar to R-KMSE, below, we present a data dependent approach to select λ using leave-one-out cross validation. While the shrinkage parameter for R-KMSE can be obtained in a simple closed form (see Proposition 9), we will see below that finding the corresponding parameter for S-KMSE is more involved. Evaluating the score function (*i.e.*, (24)) naïvely requires one to solve for $\hat{\mu}_\lambda^{(-i)}$ explicitly for every i, which is computationally expensive. The following result provides an alternate expression for the score, which can be evaluated efficiently. We would like to point out that a variation of Proposition 15 already appeared in Muandet et al. (2014a, Theorem 4). However, Theorem 4 in Muandet et al. (2014a) uses an inappropriate choice of $\hat{\mu}_\lambda^{(-i)}$, which we fixed in the following result.

Proposition 15 *The LOOCV score of S-KMSE is given by*

$$LOOCV(\lambda) = \frac{1}{n}\mathrm{tr}\left((\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{K}(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{A}_\lambda\right) - \frac{2}{n}\mathrm{tr}\left((\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{B}_\lambda\right) + \frac{1}{n}\sum_{i=1}^{n} k(x_i, x_i),$$

where $\lambda_n \triangleq (n-1)\lambda$, $\mathbf{A}_\lambda \triangleq \frac{1}{(n-1)^2}\sum_{i=1}^n \mathbf{c}_{i,\lambda}\mathbf{c}_{i,\lambda}^\top$, $\mathbf{B}_\lambda \triangleq \frac{1}{n-1}\sum_{i=1}^n \mathbf{c}_{i,\lambda}\mathbf{k}_i^\top$, $d_{i,\lambda} \triangleq \mathbf{k}_i^\top(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i$,

$$\mathbf{c}_{i,\lambda} \triangleq \mathbf{K}\mathbf{1} - \mathbf{k}_i - \mathbf{e}_i\mathbf{k}_i^\top\mathbf{1} + \mathbf{e}_i k(x_i, x_i) + \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{K}\mathbf{1}}{1 - d_{i,\lambda}} - \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{k}_i}{1 - d_{i,\lambda}} - \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top\mathbf{1}}{1 - d_{i,\lambda}} + \frac{\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n\mathbf{I})^{-1}\mathbf{e}_i k(x_i, x_i)}{1 - d_{i,\lambda}},$$

$\mathbf{k}_i$ is the i^{th} column of $\mathbf{K}$, $\mathbf{1} \triangleq (1, \ldots, 1)^\top$ and $\mathbf{e}_i \triangleq (0, 0, \ldots, 1, \ldots, 0)^\top$ with 1 being in the i^{th} position. Here $\mathrm{tr}(\mathbf{A})$ denotes the trace of a square matrix $\mathbf{A}$.

Proof See Section 5.7. ■

Unlike R-KMSE, a closed form expression for the minimizer of $LOOCV(\lambda)$ in Proposition 15 is not possible and so proving the consistency of S-KMSE along with results similar to those in Theorem 10 are highly non-trivial. Hence, we are not able to provide any theoretical comparison of $\check{\mu}_\lambda$ (with λ being chosen as a minimizer of $LOOCV(\lambda)$ in Proposition 15) with $\hat{\mu}$. However, in the next section, we provide an empirical comparison through simulations where we show that the S-KMSE outperforms the empirical estimator.

5. Proofs

In this section, we present the missing proofs of the results of Sections 2–4.

5.1 Proof of Proposition 3

($\Rightarrow$) If $\mathbb{P} = \delta_x$ for some $x \in \mathcal{X}$, then $\hat{\mu} = \mu = k(\cdot, x)$ and thus $\Delta = 0$.
($\Leftarrow$) Suppose $\Delta = 0$. It follows from (7) that $\iint (k(x,x) - k(x,y))\, \mathrm{d}\mathbb{P}(x)\, \mathrm{d}\mathbb{P}(y) = 0$. Since k is translation invariant, this reduces to

$$\iint (\psi(0) - \psi(x-y))\, \mathrm{d}\mathbb{P}(x)\, \mathrm{d}\mathbb{P}(y) = 0.$$

By invoking Bochner's theorem (Wendland, 2005, Theorem 6.6), which states that ψ is the Fourier transform of a non-negative finite Borel measure Λ, *i.e.*, $\psi(x) = \int e^{-ix^\top \omega}\, \mathrm{d}\Lambda(\omega)$, $x \in \mathbb{R}^d$, we obtain (see (16) in the proof of Proposition 5 in Sriperumbudur et al. (2011))

$$\iint \psi(x-y)\, \mathrm{d}\mathbb{P}(x)\, \mathrm{d}\mathbb{P}(y) = \int |\phi_{\mathbb{P}}(\omega)|^2\, \mathrm{d}\Lambda(\omega),$$

thereby yielding

$$\int (|\phi_{\mathbb{P}}(\omega)|^2 - 1)\, \mathrm{d}\Lambda(\omega) = 0, \tag{33}$$

where $\phi_{\mathbb{P}}$ is the characteristic function of $\mathbb{P}$. Note that $\phi_{\mathbb{P}}$ is uniformly continuous and $|\phi_{\mathbb{P}}| \leq 1$. Since k is characteristic, Theorem 9 in Sriperumbudur et al. (2010) implies that $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, using which in (33) yields $|\phi_{\mathbb{P}}(\omega)| = 1$ for all $\omega \in \mathbb{R}^d$. Since $\phi_{\mathbb{P}}$ is positive definite on $\mathbb{R}^d$, it follows from Sasvári (2013, Lemma 1.5.1) that $\phi_{\mathbb{P}}(\omega) = e^{\sqrt{-1}\omega^\top x}$ for some $x \in \mathbb{R}^d$ and thus $\mathbb{P} = \delta_x$.

5.2 Proof of Theorem 7

Before we prove Theorem 7, we present Bernstein's inequality in separable Hilbert spaces, quoted from Yurinsky (1995, Theorem 3.3.4), which will be used to prove Theorem 7.

Theorem 16 (Bernstein's inequality) *Let $(\Omega, \mathcal{A}, P)$ be a probability space, H be a separable Hilbert space, $B > 0$ and $\theta > 0$. Furthermore, let $\xi_1, \ldots, \xi_n : \Omega \to H$ be zero mean independent random variables satisfying*

$$\sum_{i=1}^n \mathbb{E}\|\xi_i\|_H^m \leq \frac{m!}{2}\theta^2 B^{m-2}. \tag{34}$$

Then for any $\tau > 0$,

$$P^n \left\{ (\xi_1, \ldots, \xi_n) : \left\| \sum_{i=1}^{n} \xi_i \right\|_H \geq 2B\tau + \sqrt{2\theta^2 \tau} \right\} \leq 2e^{-\tau}.$$

Proof (of Theorem 7) Consider

$$\begin{aligned}
\tilde{\alpha} - \alpha_* &= \frac{\hat{\Delta}}{\hat{\Delta} + \|\hat{\mu}\|^2_{\mathcal{H}}} - \frac{\Delta}{\Delta + \|\mu\|^2_{\mathcal{H}}} = \frac{\hat{\Delta}\|\mu\|^2_{\mathcal{H}} - \Delta\|\hat{\mu}\|^2_{\mathcal{H}}}{(\hat{\Delta} + \|\hat{\mu}\|^2_{\mathcal{H}})(\Delta + \|\mu\|^2_{\mathcal{H}})} \\
&= \frac{\hat{\Delta}(\|\mu\|^2_{\mathcal{H}} - \|\hat{\mu}\|^2_{\mathcal{H}})}{(\Delta + \|\mu\|^2_{\mathcal{H}})(\hat{\Delta} + \|\hat{\mu}\|^2_{\mathcal{H}})} + \frac{(\hat{\Delta} - \Delta)\|\hat{\mu}\|^2_{\mathcal{H}}}{(\Delta + \|\mu\|^2_{\mathcal{H}})(\hat{\Delta} + \|\hat{\mu}\|^2_{\mathcal{H}})} \\
&= \frac{\tilde{\alpha}(\|\mu\|^2_{\mathcal{H}} - \|\hat{\mu}\|^2_{\mathcal{H}})}{(\Delta + \|\mu\|^2_{\mathcal{H}})} + \frac{(\hat{\Delta} - \Delta)(1 - \tilde{\alpha})}{(\Delta + \|\mu\|^2_{\mathcal{H}})}.
\end{aligned}$$

Rearranging $\tilde{\alpha}$, we obtain

$$\tilde{\alpha} - \alpha_* = \frac{\alpha_*(\|\mu\|^2_{\mathcal{H}} - \|\hat{\mu}\|^2_{\mathcal{H}}) + (1 - \alpha_*)(\hat{\Delta} - \Delta)}{\hat{\Delta} + \|\hat{\mu}\|^2_{\mathcal{H}}}.$$

Therefore,

$$|\tilde{\alpha} - \alpha_*| \leq \frac{\alpha_* |\|\mu\|^2_{\mathcal{H}} - \|\hat{\mu}\|^2_{\mathcal{H}}| + (1 + \alpha_*)|\hat{\Delta} - \Delta|}{(\Delta + \|\mu\|^2_{\mathcal{H}}) - (\|\mu\|^2_{\mathcal{H}} - \|\hat{\mu}\|^2_{\mathcal{H}}) + (\hat{\Delta} - \Delta)}, \tag{35}$$

where it is easy to verify that

$$|\hat{\Delta} - \Delta| \leq \frac{|\mathbb{E}_{x,\tilde{x}} k(x, \tilde{x}) - \hat{\mathbb{E}} k(x, \tilde{x})|}{n} + \frac{|\hat{\mathbb{E}} k(x, x) - \mathbb{E}_x k(x, x)|}{n}. \tag{36}$$

In the following we obtain bounds on $|\hat{\mathbb{E}} k(x, x) - \mathbb{E}_x k(x, x)|$, $|\mathbb{E}_{x,\tilde{x}} k(x, \tilde{x}) - \hat{\mathbb{E}} k(x, \tilde{x})|$ and $|\|\mu\|^2_{\mathcal{H}} - \|\hat{\mu}\|^2_{\mathcal{H}}|$ when the kernel satisfies (17) and (18).

Bound on $|\hat{\mathbb{E}} k(x, x) - \mathbb{E}_x k(x, x)|$:

Since k is a continuous kernel on a separable topological space $\mathcal{X}$, it follows from Lemma 4.33 of Steinwart and Christmann (2008) that $\mathcal{H}$ is separable. By defining $\xi_i \triangleq k(x_i, x_i) - \mathbb{E}_x k(x.x)$, it follows from (18) that $\theta = \sqrt{n}\sigma_2$ and $B = \kappa_2$ and so by Theorem 16, for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$|\hat{\mathbb{E}} k(x, x) - \mathbb{E}_x k(x, x)| \leq \sqrt{\frac{2\sigma_2^2 \tau}{n}} + \frac{2\kappa_2 \tau}{n}. \tag{37}$$

Bound on $\|\hat{\mu} - \mu\|_{\mathcal{H}}$:

By defining $\xi_i \triangleq k(\cdot, x_i) - \mu$ and using (17), we have $\theta = \sqrt{n}\sigma_1$ and $B = \kappa_1$. Therefore, by Theorem 16, for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$\|\hat{\mu} - \mu\|_{\mathcal{H}} \leq \sqrt{\frac{2\sigma_1^2 \tau}{n}} + \frac{2\kappa_1 \tau}{n}. \tag{38}$$

Bound on $|\|\hat{\mu}\|^2_{\mathcal{H}} - \|\mu\|^2_{\mathcal{H}}|$:

Since

$$\left|\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2\right| \leq (\|\hat{\mu}\|_{\mathcal{H}} + \|\mu\|_{\mathcal{H}})\|\hat{\mu} - \mu\|_{\mathcal{H}} \leq (\|\hat{\mu} - \mu\|_{\mathcal{H}} + 2\|\mu\|_{\mathcal{H}})\|\hat{\mu} - \mu\|_{\mathcal{H}},$$

it follows from (38) that for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$\left|\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2\right| \leq D_1\sqrt{\frac{\tau}{n}} + D_2\left(\frac{\tau}{n}\right) + D_3\left(\frac{\tau}{n}\right)^{3/2} + D_4\left(\frac{\tau}{n}\right)^2, \tag{39}$$

where $(D_i)_{i=1}^4$ are positive constants that depend only on σ_1^2, κ and $\|\mu\|_{\mathcal{H}}$, and not on n and τ.

Bound on $|\hat{\mathbb{E}}k(x,\tilde{x}) - \mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})|$:

Since

$$\hat{\mathbb{E}}k(x,\tilde{x}) - \mathbb{E}_{x,\tilde{x}}k(x,\tilde{x}) = \frac{n^2(\|\hat{\mu}\|_{\mathcal{H}}^2 - \|\mu\|_{\mathcal{H}}^2) + n(\mathbb{E}_x k(x,x) - \hat{\mathbb{E}}k(x,x)) + n(\|\mu\|_{\mathcal{H}}^2 - \mathbb{E}_x k(x,x))}{n(n-1)},$$

it follows from (37) and (39) that for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$\begin{aligned}|\hat{\mathbb{E}}k(x,\tilde{x}) - \mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})| &\leq F_1\sqrt{\frac{\tau}{n}} + F_2\left(\frac{\tau}{n}\right) + F_3\left(\frac{\tau}{n}\right)^{3/2} + F_4\left(\frac{\tau}{n}\right)^2 + \frac{F_5}{n} \\ &\leq F_1'\sqrt{\frac{1+\tau}{n}} + F_2'\left(\frac{1+\tau}{n}\right) + F_3'\left(\frac{1+\tau}{n}\right)^{3/2} + F_4'\left(\frac{1+\tau}{n}\right)^2 \end{aligned} \tag{40}$$

where $(F_i)_{i=1}^5$ and $(F_i')_{i=1}^4$ are positive constants that do not depend on n and τ.

Bound on $|\tilde{\alpha} - \alpha_*|$:

Using (37) and (40) in (36), for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$|\hat{\Delta} - \Delta| \leq \frac{F_1''}{n}\sqrt{\frac{1+\tau}{n}} + \frac{F_2''}{n}\left(\frac{1+\tau}{n}\right) + \frac{F_3''}{n}\left(\frac{1+\tau}{n}\right)^{3/2} + \frac{F_4''}{n}\left(\frac{1+\tau}{n}\right)^2,$$

using which in (35) along with (39), we obtain that for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$|\tilde{\alpha} - \alpha_*| \leq \frac{\sum_{i=1}^4 \left(G_{i1}\alpha_* + \frac{G_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^4 \left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1+\tau}{n}\right)^{i/2}\right|}, \tag{41}$$

where $\theta_n \triangleq \Delta + \|\mu\|_{\mathcal{H}}^2$ and $(G_{i1})_{i=1}^4$, $(G_{i2})_{i=1}^4$ are positive constants that do not depend on n and τ. Since $\alpha_* = \frac{\Delta}{\Delta + \|\mu\|_{\mathcal{H}}^2} = \frac{\mathbb{E}_x k(x,x) - \mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})}{\mathbb{E}_x k(x,x) + (n-1)\mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})} = O(n^{-1})$ and $\theta_n = \frac{\mathbb{E}_x k(x,x) + (n-1)\|\mu\|_{\mathcal{H}}^2}{n} = O(1)$ as $n \to \infty$, it follows from (41) that $|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$.

Bound on $|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}|$:

Using (38) and (41) in

$$|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| \leq \|\hat{\mu}_{\tilde{\alpha}} - \hat{\mu}_{\alpha_*}\|_{\mathcal{H}} \leq |\tilde{\alpha} - \alpha_*|\|\hat{\mu} - \mu\|_{\mathcal{H}} + |\tilde{\alpha} - \alpha_*|\|\mu\|_{\mathcal{H}},$$

for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$, we have

$$|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| \leq \frac{\sum_{i=1}^{6} \left(G'_{i1}\alpha_* + \frac{G'_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^{4}\left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1+\tau}{n}\right)^{i/2}\right|}, \tag{42}$$

where $(G'_{i1})_{i=1}^6$ and $(G'_{i2})_{i=1}^6$ are positive constants that do not depend on n and τ. From (42), it is easy to see that $|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$.

Bound on $\mathbb{E}\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \mathbb{E}\|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2$:
Since

$$\begin{aligned}
\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 &\leq \left(\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} + \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}\right)|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| \\
&\leq 2\left(\|\hat{\mu}\|_{\mathcal{H}} + \|\mu\|_{\mathcal{H}}\right)|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| \\
&\leq 2\left(\|\hat{\mu} - \mu\|_{\mathcal{H}} + 2\|\mu\|_{\mathcal{H}}\right)|\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}|,
\end{aligned}$$

for any $\tau > 0$, with probability at least $1 - 4e^{-\tau}$,

$$\begin{aligned}
\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 &\leq \frac{\sum_{i=1}^{8}\left(G''_{i1}\alpha_* + \frac{G''_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^{4}\left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1+\tau}{n}\right)^{i/2}\right|}, \\
&\leq \frac{\sum_{i=1}^{8}\left(G''_{i1}\alpha_* + \frac{G''_{i2}}{n}(1+\alpha_*)\right)\left(\frac{1+\tau}{n}\right)^{i/2}}{\left|\theta_n - \sum_{i=1}^{4}\left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1}{n}\right)^{i/2}\right|}, \\
&\leq \begin{cases} \frac{\gamma_n}{\phi_n}\sqrt{\frac{1+\tau}{n}}, & 0 < \tau \leq n-1 \\ \frac{\gamma_n}{\phi_n}\left(\frac{1+\tau}{n}\right)^4, & \tau \geq n-1 \end{cases},
\end{aligned}$$

where $\gamma_n \triangleq H_1\alpha_* + \frac{H_2}{n}(1+\alpha_*)$, $\phi_n \triangleq \left|\theta_n - \sum_{i=1}^{4}\left(G_{i1} + \frac{G_{i2}}{n}\right)\left(\frac{1}{n}\right)^{i/2}\right|$ and $(H_i)_{i=1}^2$ are positive constants that do not depend on n and τ. In other words,

$$\mathbb{P}\left(\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 > \epsilon\right) \leq \begin{cases} 4\exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^2\right), & \frac{\gamma_n}{\phi_n\sqrt{n}} \leq \epsilon \leq \frac{\gamma_n}{\phi_n} \\ 4\exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^{1/4}\right), & \epsilon \geq \frac{\gamma_n}{\phi_n} \end{cases}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \mathbb{E}\|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 &= \int_0^\infty \mathbb{P}\left(\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 > \epsilon\right)\,\mathrm{d}\epsilon \\
&\leq \frac{\gamma_n}{\phi_n\sqrt{n}} + 4\int_{\frac{\gamma_n}{\phi_n\sqrt{n}}}^{\frac{\gamma_n}{\phi_n}} \exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^2\right)\,\mathrm{d}\epsilon \\
&\quad + 4\int_{\frac{\gamma_n}{\phi_n}}^{\infty} \exp\left(1 - n\left(\frac{\epsilon\phi_n}{\gamma_n}\right)^{1/4}\right)\,\mathrm{d}\epsilon \\
&= \frac{\gamma_n}{\phi_n\sqrt{n}} + \frac{2\gamma_n}{\phi_n\sqrt{n}}\int_0^{n-1}\frac{e^{-t}}{\sqrt{t+1}}\,\mathrm{d}t + \frac{16e\gamma_n}{n^4\phi_n}\int_n^\infty t^3 e^{-t}\,\mathrm{d}t.
\end{aligned}$$

Since $\int_0^{n-1} \frac{e^{-t}}{\sqrt{t+1}}\,\mathrm{d}t \leq \int_0^{\infty} e^{-t}\,\mathrm{d}t = 1$ and $\int_n^{\infty} t^3 e^{-t}\,\mathrm{d}t \leq \int_0^{\infty} t^3 e^{-t}\,\mathrm{d}t = 6$, we have

$$\mathbb{E}\|\hat{\mu}_{\tilde{\alpha}} - \mu\|_{\mathcal{H}}^2 - \mathbb{E}\|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}^2 \leq \frac{3\gamma_n}{\phi_n\sqrt{n}} + \frac{96e\gamma_n}{n^4\phi_n}.$$

The claim in (19) follows by noting that $\gamma_n = O(n^{-1})$ and $\phi_n = O(1)$ as $n \to \infty$. ■

5.3 Proof of Proposition 9

Define $\alpha \triangleq \frac{\lambda}{\lambda+1}$ and $\phi(x_i) \triangleq k(\cdot, x_i)$. Note that

$$\begin{aligned}
LOOCV(\lambda) &\triangleq \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{(1-\alpha)}{n-1}\sum_{j\neq i}\phi(x_j) - \phi(x_i)\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{n(1-\alpha)}{n-1}\hat{\mu} - \frac{1-\alpha}{n-1}\phi(x_i) - \phi(x_i)\right\|_{\mathcal{H}}^2 \\
&= \left\|\frac{n(1-\alpha)}{n-1}\hat{\mu}\right\|_{\mathcal{H}}^2 - \frac{2}{n}\left\langle\sum_{i=1}^{n}\frac{n-\alpha}{n-1}\phi(x_i), \frac{n(1-\alpha)}{n-1}\hat{\mu}\right\rangle_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{n-\alpha}{n-1}\phi(x_i)\right\|_{\mathcal{H}}^2 \\
&= \left(\frac{n^2(1-\alpha)^2}{(n-1)^2} - \frac{2n(n-\alpha)(1-\alpha)}{(n-1)^2}\right)\|\hat{\mu}\|_{\mathcal{H}}^2 + \frac{(n-\alpha)^2}{n(n-1)^2}\sum_{i=1}^{n}k(x_i,x_i) \\
&= \frac{1}{(n-1)^2}\left\{\alpha^2(n^2\rho - 2n\rho + \varrho) + 2n\alpha(\rho - \varrho) + n^2(\varrho - \rho)\right\} \triangleq \frac{F(\alpha)}{(n-1)^2}.
\end{aligned}$$

Since $\frac{d}{d\lambda}LOOCV(\lambda) = (n-1)^{-2}\frac{d}{d\alpha}F(\alpha)\frac{d\alpha}{d\lambda} = (n-1)^{-2}(1+\lambda)^{-2}\frac{d}{d\alpha}F(\alpha)$, equating it zero yields (25). It is easy to show that the second derivative of $LOOCV(\lambda)$ is positive implying that $LOOCV(\lambda)$ is strictly convex and so λ_r is unique.

5.4 Proof of Theorem 10

Since $\hat{\mu}_{\lambda_r} = \frac{\hat{\mu}}{1+\lambda_r} = (1-\alpha_r)\hat{\mu}$, we have $\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} \leq \alpha_r\|\hat{\mu}\|_{\mathcal{H}} + \|\hat{\mu} - \mu\|_{\mathcal{H}}$. Note that

$$\alpha_r = \frac{n(\varrho - \rho)}{n(n-2)\rho + \varrho} = \frac{n\hat{\Delta}}{\hat{\Delta} + (n-1)\|\hat{\mu}\|_{\mathcal{H}}^2} = \frac{\hat{\mathbb{E}}k(x,x) - \hat{\mathbb{E}}k(x,\tilde{x})}{\hat{\mathbb{E}}k(x,x) + (n-2)\hat{\mathbb{E}}k(x,\tilde{x})},$$

where $\hat{\Delta}$, $\|\hat{\mu}\|_{\mathcal{H}}^2$, $\hat{\mathbb{E}}k(x,x)$ and $\hat{\mathbb{E}}k(x,\tilde{x})$ are defined in Theorem 7. Consider $|\alpha_r - \alpha_*| \leq |\alpha_r - \tilde{\alpha}| + |\tilde{\alpha} - \alpha_*|$ where $\tilde{\alpha}$ is defined in (16). From Theorem 7, we have $|\tilde{\alpha} - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$ and

$$\begin{aligned}
\alpha_r - \tilde{\alpha} &= \frac{\hat{\mathbb{E}}k(x,x) - \hat{\mathbb{E}}k(x,\tilde{x})}{\hat{\mathbb{E}}k(x,x) + (n-2)\hat{\mathbb{E}}k(x,\tilde{x})} - \frac{\hat{\mathbb{E}}k(x,x) - \hat{\mathbb{E}}k(x,\tilde{x})}{2\hat{\mathbb{E}}k(x,x) + (n-2)\hat{\mathbb{E}}k(x,\tilde{x})} \\
&= \frac{\tilde{\alpha}\hat{\mathbb{E}}k(x,x)}{\hat{\mathbb{E}}k(x,x) + (n-2)\hat{\mathbb{E}}k(x,\tilde{x})} = (\tilde{\alpha} - \alpha_*)\beta + \alpha_*\beta,
\end{aligned}$$

where $\beta \triangleq \frac{\hat{\mathbb{E}}k(x,x)}{\hat{\mathbb{E}}k(x,x) + (n-2)\hat{\mathbb{E}}k(x,\tilde{x})}$. Therefore, $|\alpha_r - \tilde{\alpha}| \leq |\tilde{\alpha} - \alpha_*||\beta| + \alpha_*|\beta|$, where $\alpha_* = O(n^{-1})$ as $n \to \infty$, which follows from Remark 8(i). Since $|\hat{\mathbb{E}}k(x,x) - \mathbb{E}_x k(x,x)| = O_{\mathbb{P}}(n^{-1/2})$ and

$|\hat{\mathbb{E}}k(x,\tilde{x}) - \mathbb{E}_{x,\tilde{x}}k(x,\tilde{x})| = O_{\mathbb{P}}(n^{-1/2})$, which follow from (37) and (40) respectively, we have $|\beta| = O_{\mathbb{P}}(n^{-1})$ as $n \to \infty$. Combining the above, we have $|\alpha_r - \tilde{\alpha}| = O_{\mathbb{P}}(n^{-2})$, thereby yielding $|\alpha_r - \alpha_*| = O_{\mathbb{P}}(n^{-3/2})$. Proceeding as in Theorem 7, we have

$$|\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| \leq \|\hat{\mu}_{\lambda_r} - \mu_{\alpha_*}\|_{\mathcal{H}} \leq |\alpha_r - \alpha_*|\|\hat{\mu} - \mu\|_{\mathcal{H}} + |\alpha_r - \alpha_*|\|\mu\|_{\mathcal{H}},$$

which from the above follows that $|\|\hat{\mu}_{\lambda_r} - \mu\|_{\mathcal{H}} - \|\hat{\mu}_{\alpha_*} - \mu\|_{\mathcal{H}}| = O_{\mathbb{P}}(n^{-3/2})$ as $n \to \infty$. By arguing as in Remark 8(i), it is easy to show that $\hat{\mu}_{\lambda_r}$ is a $\sqrt{n}$-consistent estimator of μ. (27) follows by carrying out the analysis as in the proof of Theorem 7 verbatim by replacing $\tilde{\alpha}$ with α_r, while (28) follows by appealing to Remark 8(ii).

5.5 Proof of Proposition 12

First note that for any $i \in \{1, \ldots, n\}$,

$$\hat{\Sigma}_{XX}k(\cdot, x_i) = \frac{1}{n}\sum_{j=1}^{n} k(\cdot, x_j)k(x_i, x_j) = \frac{1}{n}\Phi\mathbf{k}_i^\top$$

with $\mathbf{k}_i$ being the i^{th} row of $\mathbf{K}$. This implies for any $\mathbf{a} \in \mathbb{R}^n$,

$$\hat{\Sigma}_{XX}\Phi\mathbf{a} = \hat{\Sigma}_{XX}\left(\sum_{i=1}^{n} a_i k(\cdot, x_i)\right) \overset{(*)}{=} \sum_{i=1}^{n} a_i \hat{\Sigma}_{XX}k(\cdot, x_i) = \frac{1}{n}\sum_{i=1}^{n} a_i \Phi\mathbf{k}_i^\top,$$

where $(*)$ holds since $\hat{\Sigma}_{XX}$ is a linear operator. Also, since Φ is a linear operator, we obtain

$$\hat{\Sigma}_{XX}\Phi\mathbf{a} = \frac{1}{n}\Phi\left(\sum_{i=1}^{n} a_i \mathbf{k}_i^\top\right) = \frac{1}{n}\Phi\mathbf{K}\mathbf{a}. \tag{43}$$

To prove the result, let us define $\mathbf{a} \triangleq (\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n$ and consider

$$(\hat{\Sigma}_{XX} + \lambda I)\Phi\mathbf{a} \overset{(43)}{=} n^{-1}\Phi\mathbf{K}\mathbf{a} + \lambda\Phi\mathbf{a} = \Phi(n^{-1}\mathbf{K} + \lambda\mathbf{I})\mathbf{a} = \frac{1}{n}\Phi\mathbf{K}\mathbf{1}_n \overset{(43)}{=} \hat{\Sigma}_{XX}\Phi\mathbf{1}_n = \hat{\Sigma}_{XX}\hat{\mu}.$$

Multiplying to the left on both sides of the above equation by $(\hat{\Sigma}_{XX} + \lambda I)^{-1}$, we obtain $\Phi(\mathbf{K} + n\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n = (\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}\hat{\mu}$ and the result follows by noting that $(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX} = \hat{\Sigma}_{XX}(\hat{\Sigma}_{XX} + \lambda I)^{-1}$.

5.6 Proof of Theorem 13

By Proposition 12, we have $\check{\mu}_\lambda = (\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}\hat{\mu}$. Define $\mu_\lambda \triangleq (\Sigma_{XX} + \lambda I)^{-1}\Sigma_{XX}\mu$. Let us consider the decomposition $\check{\mu}_\lambda - \mu = (\check{\mu}_\lambda - \mu_\lambda) + (\mu_\lambda - \mu)$ with

$$\begin{aligned}
\check{\mu}_\lambda - \mu_\lambda &= (\hat{\Sigma}_{XX} + \lambda I)^{-1}(\hat{\Sigma}_{XX}\hat{\mu} - \hat{\Sigma}_{XX}\mu_\lambda - \lambda\mu_\lambda) \\
&\overset{(*)}{=} (\hat{\Sigma}_{XX} + \lambda I)^{-1}(\hat{\Sigma}_{XX}\hat{\mu} - \hat{\Sigma}_{XX}\mu_\lambda - \Sigma_{XX}\mu + \Sigma_{XX}\mu_\lambda) \\
&= (\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}(\hat{\mu} - \mu) - (\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}(\mu_\lambda - \mu) \\
&\qquad + (\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX}(\mu_\lambda - \mu),
\end{aligned}$$

where we used $\lambda\mu_\lambda = \Sigma_{XX}\mu - \Sigma_{XX}\mu_\lambda$ in $(*)$. By defining $\mathcal{A}(\lambda) \triangleq \|\mu_\lambda - \mu\|_{\mathcal{H}}$, we have

$$\begin{aligned}\|\check{\mu}_\lambda - \mu\|_{\mathcal{H}} &\leq \|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}(\hat{\mu} - \mu)\|_{\mathcal{H}} + \|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}(\mu_\lambda - \mu)\|_{\mathcal{H}} \\ &\quad + \|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX}(\mu_\lambda - \mu)\|_{\mathcal{H}} + \mathcal{A}(\lambda) \\ &\leq \|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}\| \left(\|\hat{\mu} - \mu\|_{\mathcal{H}} + \mathcal{A}(\lambda)\right) + \|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX}\|\mathcal{A}(\lambda) \\ &\quad + \mathcal{A}(\lambda), \end{aligned} \tag{44}$$

where for any bounded linear operator B, $\|B\|$ denotes its operator norm. We now bound $\|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX}\|$ as follows. It is easy to show that

$$\begin{aligned}(\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX} &= \left(I - (\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \hat{\Sigma}_{XX})\right)^{-1}(\Sigma_{XX} + \lambda I)^{-1}\Sigma_{XX} \\ &= \left(\sum_{j=0}^{\infty}\left((\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \hat{\Sigma}_{XX})\right)^j\right)(\Sigma_{XX} + \lambda I)^{-1}\Sigma_{XX},\end{aligned}$$

where the last line denotes the Neumann series and therefore

$$\begin{aligned}\|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX}\| &\leq \sum_{j=0}^{\infty}\left\|(\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \hat{\Sigma}_{XX})\right\|^j \|(\Sigma_{XX} + \lambda I)^{-1}\Sigma_{XX}\| \\ &\leq \sum_{j=0}^{\infty}\left\|(\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \hat{\Sigma}_{XX})\right\|_{\mathrm{HS}}^j,\end{aligned}$$

where we used $\|(\Sigma_{XX} + \lambda I)^{-1}\Sigma_{XX}\| \leq 1$ and the fact that Σ_{XX} and $\hat{\Sigma}_{XX}$ are Hilbert-Schmidt operators on $\mathcal{H}$ as $\|\Sigma_{XX}\|_{\mathrm{HS}} \leq \kappa < \infty$ and $\|\hat{\Sigma}_{XX}\|_{\mathrm{HS}} \leq \kappa < \infty$ with κ being the bound on the kernel. Define $\eta : \mathcal{X} \to \mathrm{HS}(\mathcal{H})$, $\eta(x) = (\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \Sigma_x)$, where $\mathrm{HS}(\mathcal{H})$ is the space of Hilbert-Schmidt operators on $\mathcal{H}$ and $\Sigma_x \triangleq k(\cdot, x) \otimes k(\cdot, x)$. Observe that $\mathbb{E}\frac{1}{n}\sum_{i=1}^n \eta(x_i) = 0$. Also, for all $i \in \{1, \ldots, n\}$, $\|\eta(x_i)\|_{\mathrm{HS}} \leq \|(\Sigma_{XX} + \lambda I)^{-1}\|\|\Sigma_{XX} - \Sigma_x\|_{\mathrm{HS}} \leq \frac{2\kappa}{\lambda}$ and $\mathbb{E}\|\eta(x_i)\|_{\mathrm{HS}}^2 \leq \frac{4\kappa^2}{\lambda^2}$. Therefore, by Bernstein's inequality (see Theorem 16), for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$ over the choice of $\{x_i\}_{i=1}^n$,

$$\|(\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \hat{\Sigma}_{XX})\|_{\mathrm{HS}} \leq \frac{\kappa\sqrt{2\tau}}{\lambda\sqrt{n}} + \frac{2\kappa\tau}{\lambda n} \leq \frac{\kappa\sqrt{2\tau}(\sqrt{2\tau} + 1)}{\lambda\sqrt{n}}.$$

For $\lambda \geq \frac{\kappa\sqrt{8\tau}(\sqrt{2\tau}+1)}{\sqrt{n}}$, we obtain that $\|(\Sigma_{XX} + \lambda I)^{-1}(\Sigma_{XX} - \hat{\Sigma}_{XX})\|_{\mathrm{HS}} \leq \frac{1}{2}$ and therefore $\|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\Sigma_{XX}\| \leq 2$. Using this along with $\|(\hat{\Sigma}_{XX} + \lambda I)^{-1}\hat{\Sigma}_{XX}\| \leq 1$ and (38) in (44), we obtain that for any $\tau > 0$ and $\lambda \geq \frac{\kappa\sqrt{8\tau}(\sqrt{2\tau}+1)}{\sqrt{n}}$, with probability at least $1 - 2e^{-\tau}$ over the choice of $\{x_i\}_{i=1}^n$,

$$\|\check{\mu}_\lambda - \mu\|_{\mathcal{H}} \leq \frac{\sqrt{2\kappa\tau} + 4\tau\sqrt{\kappa}}{\sqrt{n}} + 4\mathcal{A}(\lambda). \tag{45}$$

We now analyze $\mathcal{A}(\lambda)$. Since k is continuous and $\mathcal{X}$ is separable, $\mathcal{H}$ is separable (Steinwart and Christmann, 2008, Lemma 4.33). Also Σ_{XX} is compact since it is Hilbert-Schmidt. The consistency result therefore follows from Sriperumbudur et al. (2013, Proposition A.2) which ensures $\mathcal{A}(\lambda) \to 0$ as $\lambda \to 0$. The rate also follows from Sriperumbudur et al. (2013, Proposition A.2) which yields $\mathcal{A}(\lambda) \leq \|\Sigma_{XX}^{-1}\mu\|_{\mathcal{H}}\lambda$, thereby obtaining $\|\check{\mu}_\lambda - \mu\|_{\mathcal{H}} = O_{\mathbb{P}}(n^{-1/2})$ for $\lambda = cn^{-1/2}$ with $c > 0$ being a constant independent of n.

5.7 Proof of Proposition 15

From Proposition 12, we have $\breve{\mu}_\lambda^{(-i)} = (\hat{\Sigma}^{(-i)} + \lambda I)^{-1}\hat{\Sigma}^{(-i)}\hat{\mu}^{(-i)}$ where

$$\hat{\Sigma}^{(-i)} \triangleq \frac{1}{n-1}\sum_{j\neq i} k(\cdot, x_j) \otimes k(\cdot, x_j).$$

and $\hat{\mu}^{(-i)} \triangleq \frac{1}{n-1}\sum_{j\neq i} k(\cdot, x_j)$. Define $a \triangleq k(\cdot, x_i)$. It is easy to verify that

$$\hat{\Sigma}^{(-i)} = \frac{n}{n-1}\left(\hat{\Sigma} - \frac{a\otimes a}{n}\right) \text{ and } \hat{\mu}^{(-i)} = \frac{n}{n-1}\left(\hat{\mu} - \frac{a}{n}\right).$$

Therefore,

$$\breve{\mu}_\lambda^{(-i)} = \frac{n}{n-1}\left((\hat{\Sigma} + \lambda_n' I) - \frac{a\otimes a}{n}\right)^{-1}\left(\hat{\Sigma} - \frac{a\otimes a}{n}\right)\left(\hat{\mu} - \frac{a}{n}\right),$$

which after using Sherman-Morrison formula[3] reduces to

$$\breve{\mu}_\lambda^{(-i)} = \frac{n}{n-1}\left((\hat{\Sigma} + \lambda_n' I)^{-1} + \frac{(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)(\hat{\Sigma} + \lambda_n' I)^{-1}}{n - \langle a, (\hat{\Sigma} + \lambda_n' I)^{-1} a\rangle_{\mathcal{H}}}\right)\left(\hat{\Sigma} - \frac{a\otimes a}{n}\right)\left(\hat{\mu} - \frac{a}{n}\right),$$

where $\lambda_n' \triangleq \frac{n-1}{n}\lambda$. Using the notation in the proof of Proposition 12, the following can be proved:

(i) $(\hat{\Sigma} + \lambda_n' I)^{-1}\hat{\Sigma}\hat{\mu} = n^{-1}\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{K}\mathbf{1}$.

(ii) $(\hat{\Sigma} + \lambda_n' I)^{-1}\hat{\Sigma} a = \Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{k}_i$.

(iii) $(\hat{\Sigma} + \lambda_n' I)^{-1} a = n\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i$.

Based on the above, it is easy to show that

(iv) $(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)\hat{\mu} = (\hat{\Sigma} + \lambda_n' I)^{-1} a\langle a, \hat{\mu}\rangle_{\mathcal{H}} = \Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top \mathbf{1}$.

(v) $(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a) a = (\hat{\Sigma} + \lambda_n' I)^{-1} a\langle a, a\rangle_{\mathcal{H}} = n\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i k(x_i, x_i)$.

(vi) $(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)(\hat{\Sigma} + \lambda_n' I)^{-1}\hat{\Sigma}\hat{\mu} = \Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{K}\mathbf{1}$.

(vii) $(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)(\hat{\Sigma} + \lambda_n' I)^{-1}\hat{\Sigma} a = n\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{k}_i$.

(viii) $(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)\hat{\mu} = n\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top \mathbf{1}$.

(ix) $(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a)(\hat{\Sigma} + \lambda_n' I)^{-1}(a\otimes a) a = n^2\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i\mathbf{k}_i^\top(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i k(x_i, x_i)$.

(x) $\langle a, (\hat{\Sigma} + \lambda_n' I)^{-1} a\rangle_{\mathcal{H}} = n\mathbf{k}_i^\top(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{e}_i$.

Using the above in $\breve{\mu}_\lambda^{(-i)}$, we obtain

$$\breve{\mu}_\lambda^{(-i)} = \frac{1}{n-1}\Phi(\mathbf{K} + \lambda_n \mathbf{I})^{-1}\mathbf{c}_{i,\lambda}.$$

Substituting the above in (24) yields the result.

3. The Sherman-Morrison formula states that $(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top\mathbf{A}^{-1}}{1+\mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u}}$ where $\mathbf{A}$ is an invertible square matrix, $\mathbf{u}$ and $\mathbf{v}$ are column vectors such that $1 + \mathbf{v}^\top\mathbf{A}^{-1}\mathbf{u} \neq 0$.

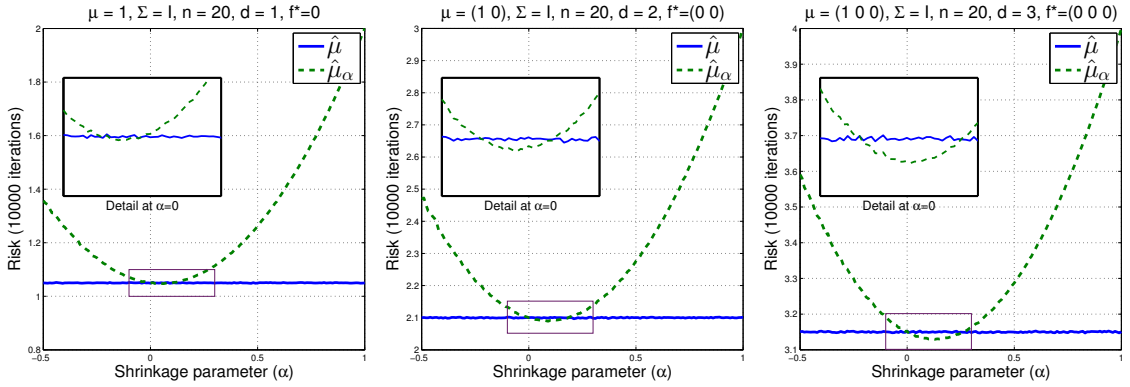


Figure 1: The comparison between standard estimator, $\hat{\mu}$ and shrinkage estimator, $\hat{\mu}_\alpha$ (with $f^* = 0$) of the mean of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^d$ where $d = 1, 2, 3$.

6. Experiments

In this section, we empirically compare the proposed shrinkage estimators to the standard estimator of the kernel mean on both synthetic and real-world datasets. Specifically, we consider the following estimators: *i*) empirical/standard kernel mean estimator (KME), *ii*) KMSE whose parameter is obtained via empirical bound (B-KMSE), *iii*) regularized KMSE whose parameter is obtained via Proposition 9 (R-KMSE), and *iv*) spectral KMSE whose parameter is obtained via Proposition 15 (S-KMSE).

6.1 Synthetic Data

Given the true data-generating distribution $\mathbb{P}$ and the i.i.d. sample $X = \{x_1, x_2, \ldots, x_n\}$ from $\mathbb{P}$, we evaluate different estimators using the loss function

$$L(\boldsymbol{\beta}, X, \mathbb{P}) \triangleq \left\| \sum_{i=1}^{n} \beta_i k(x_i, \cdot) - \mathbb{E}_{x\sim\mathbb{P}}[k(x, \cdot)] \right\|_{\mathcal{H}}^2,$$

where $\boldsymbol{\beta}$ is the weight vector associated with different estimators. Then, we can estimate the risk of the estimator by averaging over m independent copies of X, *i.e.*, $\widehat{R} = \frac{1}{m}\sum_{j=1}^{m} L(\boldsymbol{\beta}_j, X_j, \mathbb{P})$.

To allow for an exact calculation of $L(\boldsymbol{\beta}, X, \mathbb{P})$, we consider $\mathbb{P}$ to be a mixture-of-Gaussians distribution and k being one of the following kernel functions: *i*) linear kernel $k(x, x') = x^\top x'$, *ii*) polynomial degree-2 kernel $k(x, x') = (x^\top x' + 1)^2$, *iii*) polynomial degree-3 kernel $k(x, x') = (x^\top x' + 1)^3$ and *iv*) Gaussian RBF kernel $k(x, x') = \exp\left(-\|x - x'\|^2/2\sigma^2\right)$. We refer to them as LIN, POLY2, POLY3, and RBF, respectively. The analytic forms of $\mathbb{E}_{x\sim\mathbb{P}}[k(x, \cdot)]$ for Gaussian distribution are given in Song et al. (2008) and Muandet et al. (2012). Unless otherwise stated, we set the bandwidth parameter of the Gaussian kernel as $\sigma^2 = \text{median}\left\{\|x_i - x_j\|^2 : i, j = 1, \ldots, n\right\}$, *i.e.*, the median heuristic.

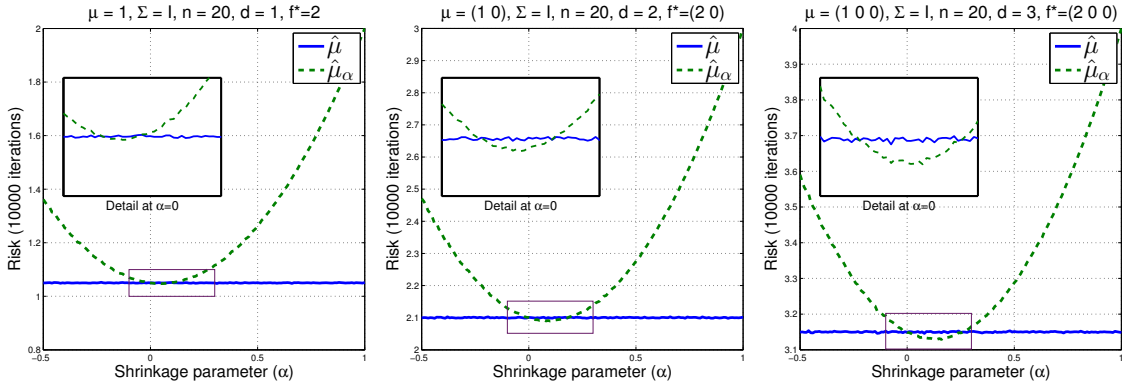


Figure 2: The risk comparison between standard estimator, $\hat{\mu}$ and shrinkage estimator, $\hat{\mu}_\alpha$ (with $f^* \in \{2, (2,0)^\top, (2,0,0)^\top\}$) of the mean of the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ on $\mathbb{R}^d$ where $d = 1, 2, 3$.

6.1.1 Gaussian Distribution

We begin our empirical studies by considering the simplest case in which the distribution $\mathbb{P}$ is a Gaussian distribution $\mathcal{N}(\mu, \mathbf{I})$ on $\mathbb{R}^d$ where $d = 1, 2, 3$ and k is a linear kernel. In this case, the problem of kernel mean estimation reduces to just estimating the mean μ of the Gaussian distribution $\mathcal{N}(\mu, \mathbf{I})$. We consider only shrinkage estimators of form $\hat{\mu}_\alpha = \alpha f^* + (1-\alpha)\hat{\mu}$. The true mean μ of the distribution is chosen to be 1, $(1,0)^\top$, and $(1,0,0)^\top$, respectively. Figure 1 depicts the comparison between the standard estimator and the shrinkage estimator, $\hat{\mu}_\alpha$ when the target f^* is the origin. We can clearly see that even in this simple case, an improvement can be gained by applying a small shrinkage. Furthermore, the improvement becomes more substantial as we increase the dimensionality of the underlying space. Figure 2 illustrates similar results when $f^* \neq 0$ but $f^* \in \{2, (2,0)^\top, (2,0,0)^\top\}$. Interestingly, we can still observe similar improvement, which demonstrates that the choice of target f^* can be arbitrary when no prior knowledge about $\mu_{\mathbb{P}}$ is available.

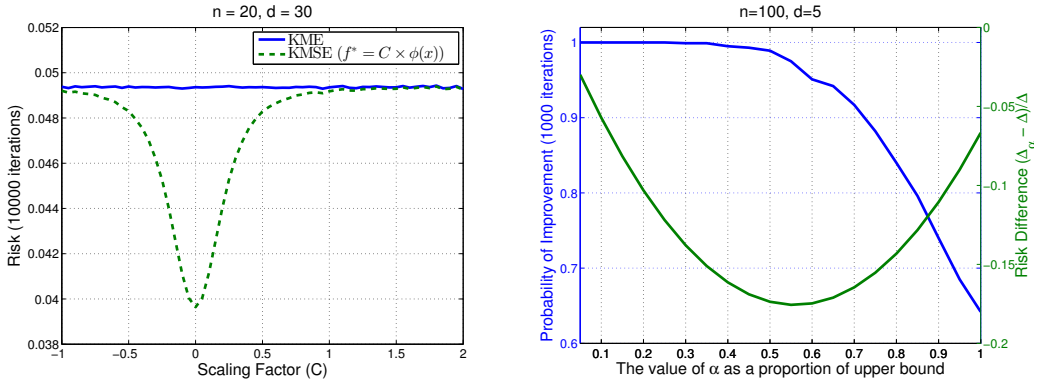
6.1.2 Mixture of Gaussians Distributions

To simulate a more realistic case, let y be a sample from $\mathbb{P} \triangleq \sum_{i=1}^4 \pi_i \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_i)$. In the following experiments, the sample x is generated from the following generative process:

$$x = y + \varepsilon, \quad \theta_{ij} \sim \mathcal{U}(-10, 10), \quad \Sigma_i \sim \mathcal{W}(2 \times \mathbf{I}_d, 7), \quad \varepsilon \sim \mathcal{N}(0, 0.2 \times \mathbf{I}_d),$$

where $\mathcal{U}(a, b)$ and $\mathcal{W}(\Sigma_0, df)$ represent the uniform distribution and Wishart distribution, respectively. We set $\boldsymbol{\pi} = (0.05, 0.3, 0.4, 0.25)^\top$. The choice of parameters here is quite arbitrary; we have experimented using various parameter settings and the results are similar to those presented here.

Figure 3(a) depicts the comparison between the standard kernel mean estimator and the shrinkage estimator, $\hat{\mu}_\alpha$ when the kernel k is the Gaussian RBF kernel. For shrinkage estimator $\hat{\mu}_\alpha$, we consider $f^* = C \times k(x, \cdot)$ where C is a scaling factor and each element of x is a realization of uniform random variable on $(0, 1)$. That is, we allow the target f^* to change



(a) (b)

Figure 3: (a) The risk comparison between $\hat{\mu}$ (KME) and $\hat{\mu}_{\tilde{\alpha}}$ (KMSE) where $\tilde{\alpha} = \hat{\Delta}/(\hat{\Delta} + \|f^* - \hat{\mu}\|^2_{\mathcal{H}})$. We consider when $f^* = C \times k(x, \cdot)$ where x is drawn uniformly from a pre-specified range and C is a scaling factor. (b) The probability of improvement and the risk difference as a function of shrinkage parameter α averaged over 1,000 iterations. As the value of α increases, we get more improvement in term of the risk, whereas the probability of improvement decreases as a function of α.

depending on the value of C. As the absolute value of C increases, the target function f^* will move further away from the origin. The shrinkage parameter α is determined using the empirical bound, *i.e.*, $\tilde{\alpha} = \hat{\Delta}/(\hat{\Delta} + \|f^* - \hat{\mu}\|^2_{\mathcal{H}})$. As we can see in Figure 3(a), the results reveal how important the choice of f^* is. That is, we may get substantial improvement over the empirical estimator if appropriate prior knowledge is incorporated through f^*, which in this case suggests that f^* should lie close to the origin. We intend to investigate the topic of prior knowledge in more detail in our future work.

Previous comparisons between standard estimator and shrinkage estimator is based entirely on the notion of a risk, which is in fact not useful in practice as we only observe a single copy of sample from the probability distribution. Instead, one should also look at the probability that, given a single copy of sample, the shrinkage estimator outperforms the standard one in term of a loss. To this end, we conduct an experiment comparing the standard estimator and shrinkage estimator using the Gaussian RBF kernel. In addition to the risk comparison, we also compare the probability that the shrinkage estimator gives smaller loss than that of the standard estimator. To be more precise, the probability is defined as a proportion of the samples drawn from the same distribution whose shrinkage loss is smaller than the loss of the standard estimator. Figure 3(b) illustrates the risk difference $(\Delta_\alpha - \Delta)$ and the probability of improvement (*i.e.*, the fraction of times $\Delta_\alpha < \Delta$) as a function of shrinkage parameter α. In this case, the value of α is specified as a proportion of empirical upper bound $2\hat{\Delta}/(\hat{\Delta} + \|\hat{\mu}\|^2_{\mathcal{H}})$. The results suggest that the shrinkage parameter α controls the trade-off between the amount of improvement in terms of risk and the probability that the shrinkage estimator will improve upon the standard one. However, this trade-off only holds up to a certain value of α. As α becomes too large, both the probability of improvement and the amount of improvement itself decrease, which coincides with the intuition given for the positive-part shrinkage estimators (cf. Section 2.2.1).

6.1.3 Shrinkage Estimators via Leave-One-Out Cross-Validation

In addition to the empirical upper bound, one can alternatively compute the shrinkage parameter using leave-one-out cross-validation proposed in Section 3. Our goal here is to compare the B-KMSE, R-KMSE and S-KMSE on synthetic data when the shrinkage parameter λ is chosen via leave-one-out cross-validation procedure. Note that the only difference between B-KMSE and R-KMSE is the way we compute the shrinkage parameter.

Figure 4 shows the empirical risk of different estimators using different kernels as we increase the value of shrinkage parameter λ (note that R-KMSE and S-KMSE in Figure 4 refer to those in (22) and (31) respectively). Here we scale the shrinkage parameter by the smallest non-zero eigenvalue γ_0 of the kernel matrix $\mathbf{K}$. In general, we find that R-KMSE and S-KMSE outperforms KME. Nevertheless, as the shrinkage parameter λ becomes large, there is a tendency that the specific shrinkage estimate might actually perform worse than the KME, *e.g.*, see LIN kernel and outliers in Figure 4. The result also supports our previous observation regarding Figure 3(b), which suggests that it is very important to choose the parameter λ appropriately.

To demonstrate the leave-one-out cross-validation procedure, we conduct similar experiments in which the parameter λ is chosen by the proposed LOOCV procedure. Figure 5 depicts the percentage of improvement (with respect to the empirical risk of the KME[4]) as we vary the sample size and dimension of the data. Clearly, B-KMSE, R-KMSE and S-KMSE outperform the standard estimator. Moreover, both R-KMSE and S-KMSE tend to outperform the B-KMSE. We can also see that the performance of S-KMSE depends on the choice of kernel. This makes sense intuitively because S-KMSE also incorporates the eigen-spectrum of $\mathbf{K}$, whereas R-KMSE does not. The effects of both sample size and data dimensionality are also transparent from Figure 5. While it is intuitive to see that the improvement gets smaller with increase in sample size, it is a bit surprising to see that we can gain much more in high-dimensional input space, especially when the kernel function is non-linear, because the estimation happens in the feature space associated with the kernel function rather than in the input space. Lastly, we note that the improvement is more substantial in the "large d, small n" paradigm.

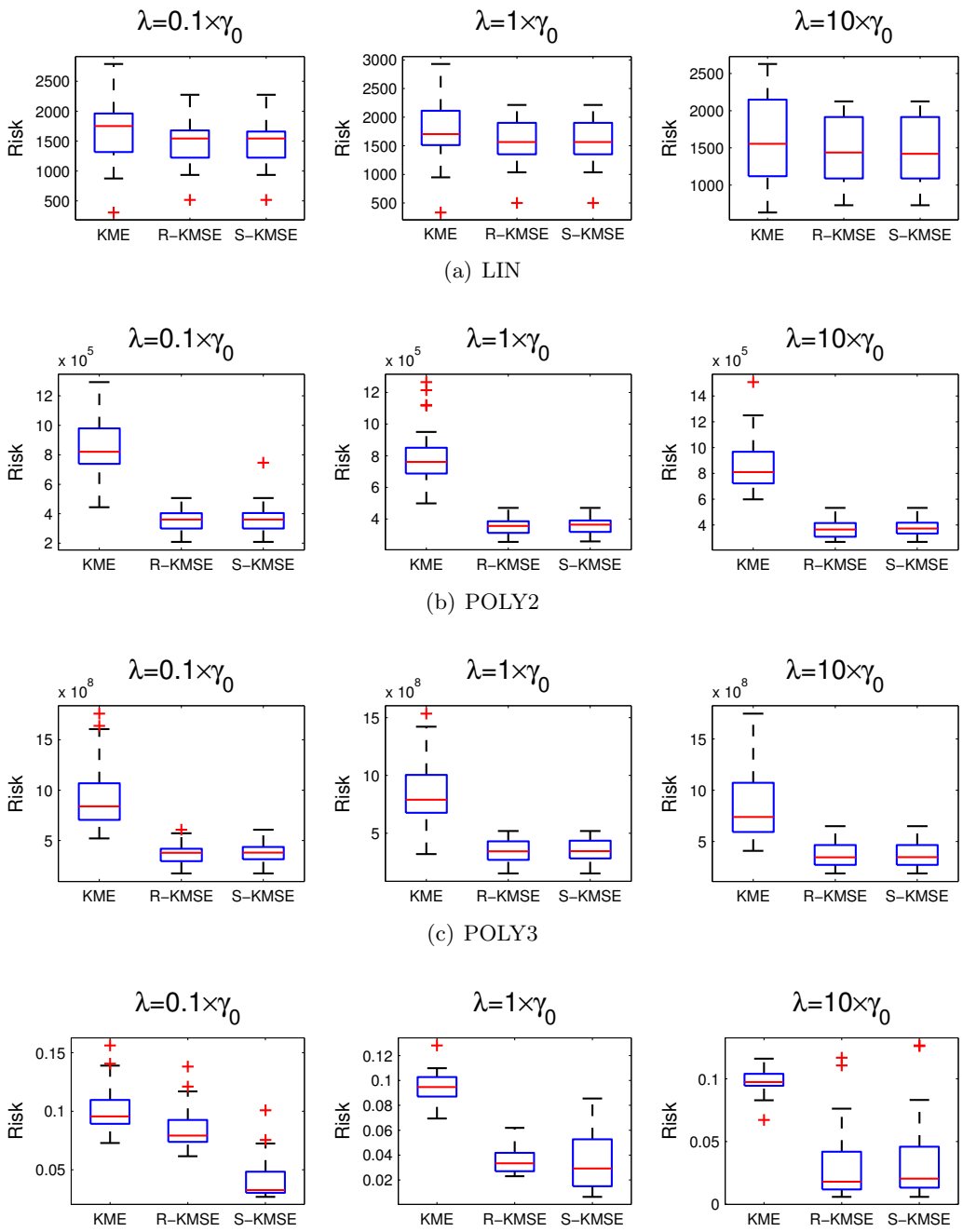
6.2 Real Data

To evaluate the proposed estimators on real-world data, we consider several benchmark applications, namely, classification via Parzen window classifier, density estimation via kernel mean matching (Song et al., 2008), and discriminative learning on distributions (Muandet et al., 2012; Muandet and Schölkopf, 2013). For some of these tasks we employ datasets from the UCI repositories. We use only real-valued features, each of which is normalized to have zero mean and unit variance.

6.2.1 Parzen Window Classifiers

One of the oldest and best-known classification algorithms is the *Parzen window classifier* (Duda et al., 2000). It is easy to implement and is one of the powerful non-linear supervised

4. If we denote the loss of KME and KMSE as ℓ_{KME} and ℓ_{KMSE}, respectively, the percentage of improvement is calculated as $100 \times (\ell_{KME} - \ell_{KMSE})/\ell_{KME}$.



(a) LIN

(b) POLY2

(c) POLY3

(d) RBF

Figure 4: The average loss of KME (left), R-KMSE (middle) and S-KMSE (right) estimators with different values of shrinkage parameter. We repeat the experiments over 30 different distributions with $n = 10$ and $d = 30$.

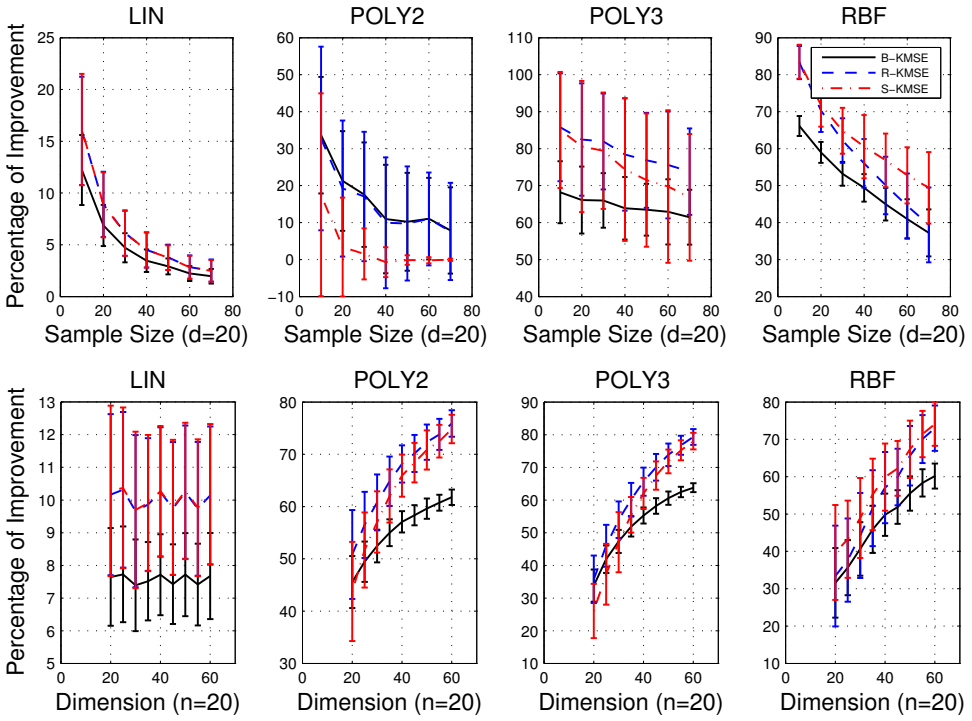


Figure 5: The percentage of improvement compared to KME over 30 different distributions of B-KMSE, R-KMSE and S-KMSE with varying sample size (n) and dimension (d). For B-KMSE, we calculate α using (16), whereas R-KMSE and S-KMSE use LOOCV to choose λ.

learning techniques. Suppose we have data points from two classes, namely, positive class and negative class. For positive class, we observe $\mathfrak{X} \triangleq \{x_1, x_2, \ldots, x_n\} \subset \mathcal{X}$, while for negative class we have $\mathfrak{Y} \triangleq \{y_1, y_2, \ldots, y_m\} \subset \mathcal{X}$. Following Shawe-Taylor and Cristianini (2004, Sec. 5.1.2), the Parzen window classifier is given by

$$f(z) = \operatorname{sgn}\left(\frac{1}{n}\sum_{i=1}^{n} k(z, x_i) - \frac{1}{m}\sum_{j=1}^{m} k(z, y_j) + b\right) = \operatorname{sgn}\left(\hat{\mu}_{\mathfrak{X}}(z) - \hat{\mu}_{\mathfrak{Y}}(z) + b\right), \tag{46}$$

where b is a bias term given by $b = \frac{1}{2}(\|\hat{\mu}_{\mathfrak{Y}}\|_{\mathcal{H}}^2 - \|\hat{\mu}_{\mathfrak{X}}\|_{\mathcal{H}}^2)$. Note that $f(z)$ is a threshold linear function in $\mathcal{H}$ with weight vector $\mathbf{w} = (1/n)\sum_{i=1}^{n}\phi(x_i) - (1/m)\sum_{j=1}^{m}\phi(y_j)$ (see Shawe-Taylor and Cristianini (2004, Sec. 5.1.2) for more detail). This algorithm is often referred to as the lazy algorithm as it does not require training.

In brief, the classifier (46) assigns the data point z to the class whose empirical kernel mean $\hat{\mu}$ is closer to the feature map $k(z, \cdot)$ of the data point in the RKHS. On the other hand, we may view the empirical kernel mean $\hat{\mu}_{\mathfrak{X}} \triangleq \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot)$ (resp. $\hat{\mu}_{\mathfrak{Y}} \triangleq \frac{1}{m}\sum_{j=1}^{m} k(y_j, \cdot)$) as a standard empirical estimate, *i.e.*, KME, of the true kernel mean representation of the class-conditional distribution $\mathbb{P}(X|Y = +1)$ (resp. $\mathbb{P}(X|Y = -1)$). Given the improvement of shrinkage estimators over the empirical estimator of kernel mean, it is natural to expect

Dataset	Classification Error Rate			
	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 1: 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.

that the performance of Parzen window classifier can be improved by employing shrinkage estimators of the true mean representation.

Our goal in this experiment is to compare the performance of Parzen window classifier using different kernel mean estimators. That is, we replace $\hat{\mu}_{\mathfrak{X}}$ and $\hat{\mu}_{\mathfrak{Y}}$ by their shrinkage counterparts and evaluate the resulting classifiers across several datasets taken from the UCI machine learning repository. In this experiment, we only consider the Gaussian RBF kernel whose bandwidth parameter is chosen by cross-validation procedure over a uniform grid $\sigma \in [0.1, 2]$. We use 30% of each dataset as a test set and the rest as a training set. We employ a simple pairwise coupling and majority vote for multi-class classification. We repeat the experiments 100 times and perform the paired-sample t-test on the results at 5% significance level. Table 1 reports the classification error rates of the Parzen window classifiers with different kernel mean estimators. Although the improvement is not substantial, we can see that the shrinkage estimators consistently give better performance than the standard estimator.

6.2.2 Density Estimation

We perform density estimation via kernel mean matching (Song et al., 2008), wherein we fit the density $Q = \sum_{j=1}^{m} \pi_j \mathcal{N}(\boldsymbol{\theta}_j, \sigma_j^2 \mathbf{I})$ to each dataset by the following minimization problem:

$$\min_{\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\sigma}} \|\hat{\mu} - \mu_Q\|_{\mathcal{H}}^2 \quad \text{subject to} \quad \sum_{j=1}^{m} \pi_j = 1,\ \pi_j \geq 0\,. \tag{47}$$

The empirical mean map $\hat{\mu}$ is obtained from samples using different estimators, whereas μ_Q is the kernel mean embedding of the density Q. Unlike experiments in Song et al. (2008), our goal is to compare different estimators of $\mu_{\mathbb{P}}$ (where $\mathbb{P}$ is the true data distribution), by replacing $\hat{\mu}$ in (47) with different shrinkage estimators. A better estimate of $\mu_{\mathbb{P}}$ should lead to better density estimation, as measured by the negative log-likelihood of Q on the test set, which we choose to be 30% of the dataset. For each dataset, we set the number of mixture components m to be 10. The model is initialized by running 50 random initializations using

the k-means algorithm and returning the best. We repeat the experiments 30 times and perform the paired sign test on the results at 5% significance level.[5]

The average negative log-likelihood of the model Q, optimized via different estimators, is reported in Table 2. In most cases, both R-KMSE and S-KMSE consistently achieve smaller negative log-likelihood when compared to KME. B-KMSE also tends to outperform the KME. However, in few cases the KMSEs achieve larger negative log-likelihood, especially when we use linear and degree-2 polynomial kernels. This highlight the potential of our estimators in a non-linear setting.

6.2.3 Discriminative Learning on Probability Distributions

The last experiment involves the discriminative learning on a collection of probability distributions via the kernel mean representation. A positive semi-definite kernel between distributions can be defined via their kernel mean embeddings. That is, given a training sample $(\widehat{\mathbb{P}}_1, y_1), \ldots, (\widehat{\mathbb{P}}_m, y_m) \in \mathcal{P} \times \{-1, +1\}$ where $\widehat{\mathbb{P}}_i := \frac{1}{n_i} \sum_{p=1}^{n_i} \delta_{x_p^i}$ and $x_p^i \sim \mathbb{P}_i$, the linear kernel between two distributions is approximated by

$$\langle \hat{\mu}_{\mathbb{P}_i}, \hat{\mu}_{\mathbb{P}_j} \rangle_{\mathcal{H}} = \left\langle \sum_{p=1}^{n_i} \beta_p^i \phi(x_p^i), \sum_{q=1}^{n_j} \beta_q^j \phi(x_q^j) \right\rangle_{\mathcal{H}} = \sum_{p=1}^{n_i} \sum_{q=1}^{n_j} \beta_p^i \beta_q^j k(x_p^i, x_q^j),$$

where the weight vectors $\boldsymbol{\beta}^i$ and $\boldsymbol{\beta}^j$ come from the kernel mean estimates of $\mu_{\mathbb{P}_i}$ and $\mu_{\mathbb{P}_j}$, respectively. The non-linear kernel can then be defined accordingly, *e.g.*, $\kappa(\mathbb{P}_i, \mathbb{P}_j) = \exp(\|\hat{\mu}_{\mathbb{P}_i} - \hat{\mu}_{\mathbb{P}_j}\|_{\mathcal{H}}^2 / 2\sigma^2)$, see Christmann and Steinwart (2010). Our goal in this experiment is to investigate if the shrinkage estimators of the kernel mean improve the performance of discriminative learning on distributions. To this end, we conduct experiments on natural scene categorization using support measure machine (SMM) (Muandet et al., 2012) and group anomaly detection on a high-energy physics dataset using one-class SMM (OCSMM) (Muandet and Schölkopf, 2013). We use both linear and non-linear kernels where the Gaussian RBF kernel is employed as an embedding kernel (Muandet et al., 2012). All hyper-parameters are chosen by 10-fold cross-validation.[6] For our unsupervised problem, we repeat the experiments using several parameter settings and report the best results. Table 3 reports the classification accuracy of SMM and the area under ROC curve (AUC) of OCSMM using different kernel mean estimators. All shrinkage estimators consistently lead to better performance on both SMM and OCSMM when compared to KME.

In summary, the proposed shrinkage estimators outperform the standard KME. While B-KMSE and R-KMSE are very competitive compared to KME, S-KMSE tends to outperform both B-KMSE and R-KMSE, however, sometimes leading to poor estimates depending on the dataset and the kernel function.

5. The paired sign test is a nonparametric test that can be used to examine whether two paired samples have the same distribution. In our case, we compare B-KMSE, R-KMSE and S-KMSE against KME.
6. In principle one can incorporate the shrinkage parameter into the cross-validation procedure. In this work we are only interested in the value of λ returned by the proposed LOOCV procedure.

Dataset	LIN				POLY2				POLY3				RBF			
	KME	B-KMSE	R-KMSE	S-KMSE	KME	B-KMSE	R-KMSE	S-KMSE	KME	B-KMSE	R-KMSE	S-KMSE	KME	B-KMSE	R-KMSE	S-KMSE
1. ionosphere	39.878	40.038	39.859	39.823	34.651	**34.352**	34.390	**34.009**	35.943	**35.575**	35.543	**34.617**	41.601	40.976	**40.817**	41.229
2. sonar	72.240	**72.044**	72.198	**72.157**	100.420	**99.573**	**97.844**	**97.783**	72.294	**71.933**	72.003	71.835	98.540	**95.815**	**93.458**	**93.010**
3. Australian	18.277	18.280	18.294	18.293	18.357	18.381	18.391	18.429	18.611	18.463	18.466	18.495	19.428	19.325	19.418	19.393
4. specft	57.444	**57.2808**	57.218	**57.224**	67.018	66.979	**66.431**	**66.391**	59.585	**58.969**	60.006	60.616	65.674	65.138	65.039	**64.699**
5. wdbc	31.801	31.759	31.776	31.781	32.421	32.310	32.373	32.316	31.183	**31.167**	**31.127**	31.110	36.471	36.453	36.335	35.898
6. wine	16.019	16.000	16.039	16.009	17.070	16.920	**16.886**	16.960	16.393	16.300	16.309	16.202	17.569	17.546	17.498	17.498
7. satimage	25.258	25.317	25.219	25.186	24.214	24.111	24.132	24.259	25.284	25.276	25.239	25.263	23.741	23.753	23.728	**24.384**
8. segment	18.326	**17.868**	18.055	**18.124**	18.571	18.292	18.277	18.631	19.642	19.549	19.404	19.628	21.946	**21.598**	**21.580**	**21.822**
9. vehicle	16.633	16.519	16.521	16.499	16.096	**15.998**	16.031	16.041	16.288	16.278	**16.281**	**16.263**	18.260	18.056	18.119	**17.911**
10. svmguide2	27.298	27.273	27.281	27.276	27.812	**28.030**	27.985	**27.975**	28.014	**28.177**	**28.321**	28.250	28.132	28.122	28.119	28.020
11. vowel	12.632	12.626	12.629	12.656	12.532	12.471	12.479	12.472	13.069	13.061	13.056	**13.054**	13.526	13.486	13.462	**13.453**
12. housing	14.637	14.441	14.469	**14.296**	15.543	15.467	15.414	15.390	15.592	**15.543**	**15.509**	**15.408**	16.487	16.239	16.424	**16.019**
13. bodyfat	17.527	17.362	**17.348**	17.396	17.386	17.358	17.356	17.329	16.418	16.393	**16.305**	**16.194**	17.875	17.652	17.607	**17.651**
14. abalone	5.706	5.665	5.708	5.722	7.281	7.116	7.185	7.025	5.864	5.847	5.853	5.832	6.068	6.039	6.049	5.910
15. glass	9.245	9.211	9.198	9.217	8.571	8.473	8.457	8.414	9.050	8.991	9.012	**8.737**	9.606	**9.605**	**9.575**	**9.573**

Table 2: Average negative log-likelihood of the model Q on test points over 30 randomizations. The boldface represents the result whose difference from the baseline, *i.e.*, KME, is statistically significant.

Estimator	**Linear Kernel**		**Non-linear Kernel**	
	SMM	OCSMM	SMM	OCSMM
KME	0.5432	0.6955	0.6017	0.9085
B-KMSE	0.5455	0.6964	0.6106	0.9088
R-KMSE	0.5521	0.6970	0.6303	0.9105
S-KMSE	0.5606	0.6970	0.6412	0.9063

Table 3: The classification accuracy of SMM and the area under ROC curve (AUC) of OCSMM using different estimators to construct the kernel on distributions.

7. Conclusion and Discussion

Motivated by the classical James-Stein phenomenon, in this paper, we proposed a shrinkage estimator for the kernel mean μ in a reproducing kernel Hilbert space $\mathcal{H}$ and showed they improve upon the empirical estimator $\hat{\mu}$ in the mean squared sense. We showed the proposed shrinkage estimator $\tilde{\mu}$ (with the shrinkage parameter being learned from data) to be $\sqrt{n}$-consistent and satisfies $\mathbb{E}\|\tilde{\mu} - \mu\|_{\mathcal{H}}^2 < \mathbb{E}\|\hat{\mu} - \mu\|_{\mathcal{H}}^2 + O(n^{-3/2})$ as $n \to \infty$. We also provided a regularization interpretation to shrinkage estimation, using which we also presented two shrinkage estimators, namely regularized shrinkage estimator and spectral shrinkage estimator, wherein the first one is closely related to $\tilde{\mu}$ while the latter exploits the spectral decay of the covariance operator in $\mathcal{H}$. We showed through numerical experiments that the proposed estimators outperform the empirical estimator in various scenarios. Most importantly, the shrinkage estimators not only provide more accurate estimation, but also lead to superior performance on many real-world applications.

In this work, while we focused mainly on an estimation of the mean function in RKHS, it is quite straightforward to extend the shrinkage idea to estimate covariance (and cross-covariance) operators and tensors in RKHS (see Appendix A for a brief description). The key observation is that the covariance operator can be viewed as a mean function in a tensor RKHS. Covariance operators in RKHS are ubiquitous in many classical learning algorithms such as kernel PCA, kernel FDA, and kernel CCA. Recently, a preliminary investigation with some numerical results on shrinkage estimation of covariance operators is carried out in Muandet et al. (2014a) and Wehbe and Ramdas (2015). In the future, we intend to carry out a detailed study on the shrinkage estimation of covariance (and cross-covariance) operators.

Acknowledgments

The authors thanks the reviewers and the action editor for their detailed comments that signficantly improved the manuscript. This work was partly done while Krikamol Muandet was visiting the Institute of Statistical Mathematics, Tokyo, and New York University, New York; and while Bharath Sriperumbudur was visiting the Max Planck Institute for Intelligent Systems, Germany. The authors wish to thank David Hogg and Ross Fedely for reading the first draft and giving valuable comments. We also thank Motonobu Kanagawa, Yu Nishiyama, and Ingo Steinwart for fruitful discussions. Kenji Fukumizu has been supported in part by MEXT Grant-in-Aid for Scientific Research on Innovative Areas 25120012.

Appendix A. Shrinkage Estimation of Covariance Operator

Let $(\mathcal{H}_X, k_X)$ and $(\mathcal{H}_Y, k_Y)$ be separable RKHSs of functions on measurable spaces $\mathcal{X}$ and $\mathcal{Y}$, with measurable reproducing kernels k_X and k_Y (with corresponding feature maps ϕ and φ), respectively. We consider a random vector $(X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$ with distribution $\mathbb{P}_{XY}$. The marginal distributions of X and Y are denoted by $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. If $\mathbb{E}_X k_X(X, X) < \infty$ and $\mathbb{E}_Y k_Y(Y, Y) < \infty$, then there exists a unique *cross-covariance*

operator $\Sigma_{YX} : \mathcal{H}_X \to \mathcal{H}_Y$ such that

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_Y} = \mathbb{E}_{XY}[(f(X) - \mathbb{E}_X[f(X)])(g(Y) - \mathbb{E}_Y[g(Y)])] = Cov(f(X), g(Y))$$

holds for all $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$ (Baker, 1973; Fukumizu et al., 2004). If X is equal to Y, we obtain the self-adjoint operator Σ_{XX} called the *covariance operator*. Given i.i.d sample $\{(x_i, y_i)\}_{i=1}^n$ from $\mathbb{P}_{XY}$, we can write the empirical cross-covariance operator $\widehat{\Sigma}_{YX}$ as

$$\widehat{\Sigma}_{YX} \triangleq \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \varphi(y_i) - \hat{\mu}_X \otimes \hat{\mu}_Y \tag{48}$$

where $\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n \phi(x_i)$ and $\hat{\mu}_Y = \frac{1}{n} \sum_{i=1}^n \varphi(y_i)$.[7] Let $\tilde{\phi}$ and $\tilde{\varphi}$ be the centered version of the feature map ϕ and φ defined as $\tilde{\phi}(x) = \phi(x) - \hat{\mu}_X$ and $\tilde{\varphi}(y) = \varphi(y) - \hat{\mu}_Y$, respectively. Then, the empirical cross-covariance operator in (48) can be rewritten as

$$\widehat{\Sigma}_{YX} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\phi}(x_i) \otimes \tilde{\varphi}(y_i),$$

and therefore a shrinkage estimator of Σ_{YX}, *e.g.*, an equivalent of B-KMSE, can be constructed based on the ideas presented in this paper. That is, by the inner product property in product space, we have

$$\begin{aligned} \langle \tilde{\phi}(x) \otimes \tilde{\varphi}(y), \tilde{\phi}(x') \otimes \tilde{\varphi}(y') \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} &= \langle \tilde{\phi}(x), \tilde{\phi}(x') \rangle_{\mathcal{H}_X} \langle \tilde{\varphi(y)}, \tilde{\varphi(y')} \rangle_{\mathcal{H}_Y} \\ &= \tilde{k}_X(x, x') \tilde{k}_Y(y, y'). \end{aligned}$$

where $\tilde{k}_X$ and $\tilde{k}_Y$ denote the centered kernel functions. As a result, we can obtain the shrinkage estimators for Σ_{YX} by plugging the above kernel into the KMSEs.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Charles R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:pp. 273–289, 1973.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52 – 72, 2007. ISSN 0885-064X.

James Berger and Robert Wolpert. Estimating the mean function of a Gaussian process and the Stein effect. *Journal of Multivariate Analysis*, 13(3):401–424, 1983.

James O. Berger. Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. *Annals of Statistics*, 4(1):223–226, 1976.

7. Although it is possible to estimate $\hat{\mu}_X$ and $\hat{\mu}_Y$ using our shrinkage estimators, the key novelty here is to directly shrink the *centered* covariance operator.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

Andreas Christmann and Ingo Steinwart. Universal kernels on Non-Standard input spaces. In *Advances in Neural Information Processing Systems (NIPS)*, pages 406–414. 2010.

Inderjit S. Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 551–556, New York, NY, USA, 2004.

Joseph Diestel and John J. Uhl. *Vector Measures*. American Mathematical Society, Providence, 1977.

Nicolae Dinculeanu. *Vector Integration and Stochastic Integration in Banach Spaces*. Wiley, 2000.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

Bradley Efron and Carl N. Morris. Stein's paradox in statistics. *Scientific American*, 236 (5):119–127, 1977.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004.

Kenji Fukumizu, Francis R. Bach, and Arthur Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383, 2007.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1737–1745. 2011.

Arthur Gretton, Karsten M. Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alexander J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, pages 585–592. MIT Press, 2008.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Marvin Gruber. *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. Statistics Textbooks and Monographs. Marcel Dekker, 1998.

Steffen Grünewälder, Guy Lever, Arthur Gretton, Luca Baldassarre, Sam Patterson, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Steffen Grünewälder, Arthur Gretton, and John Shawe-Taylor. Smooth operators. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

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

JooSeuk Kim and Clayton D. Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13:2529–2565, Sep 2012.

Avi Mandelbaum and L. A. Shepp. Admissibility as a touchstone. *Annals of Statistics*, 15 (1):252–268, 1987.

Krikamol Muandet and Bernhard Schölkopf. One-class support measure machines for group anomaly detection. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2013.

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems (NIPS)*, pages 10–18. 2012.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. Kernel mean estimation and Stein effect. In *ICML*, pages 10–18, 2014a.

Krikamol Muandet, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean estimation via spectral filtering. In *Advances in Neural Information Processing Systems 27*, pages 1–9. Curran Associates, Inc., 2014b.

Nicolas Privault and Anthony Rveillac. Stein estimation for the drift of Gaussian processes using the Malliavin calculus. *Annals of Statistics*, 36(5):2531–2550, 2008.

Carl E. Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Michael Reed and Barry Simon. *Functional Analysis*. Academic Press, New York, 1972.

Zoltán Sasvári. *Multivariate Characteristic and Correlation Functions*. De Gruyter, Berlin, Germany, 2013.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, pages 416–426, London, UK, UK, 2001. Springer-Verlag.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK, 2004.

Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT)*, pages 13–31. Springer-Verlag, 2007.

Le Song, Xinhua Zhang, Alex Smola, Arthur Gretton, and Bernhard Schölkopf. Tailoring density estimation via reproducing kernel moment matching. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 992–999, 2008.

Le Song, Byron Boots, Sajid M. Siddiqi, Geoffrey Gordon, and Alexander J. Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

Le Song, Ankur P. Parikh, and Eric P. Xing. Kernel embeddings of latent tree graphical models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2708–2716, 2011.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *The 21st Annual Conference on Learning Theory (COLT)*, 2008.

Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

Bharath Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12: 2389–2410, 2011.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. doi: 10.1214/12-EJS722.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. 2013. http://arxiv.org/pdf/1312.3516.

Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206. University of California Press, 1955.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, New York, 2008.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Larry Wasserman. *All of Nonparametric Statistics*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Leila Wehbe and Aaditya Ramdas. Nonparametric independence testing for small sample sizes. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3777–3783, July 2015.

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

Vadim Yurinsky. *Sums and Gaussian Vectors*, volume 1617 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.