Kernel Mean Estimation and Stein Effect

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

Kenji Fukumizu FUKUMIZU@ISM.AC.JP
The Institute of Statistical Mathematics, Tokyo, Japan

Bharath Sriperumbudur BS493@STATSLAB.CAM.AC.UK
Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit, University College London, London, United Kingdom

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

Abstract

A mean function in a reproducing kernel Hilbert space (RKHS), or a kernel mean, is an important part of many algorithms ranging from kernel principal component analysis to Hilbert-space embedding of distributions. Given a finite sample, an empirical average is the standard estimate for the true kernel mean. We show that this estimator can be improved due to a well-known phenomenon in statistics called Stein's phenomenon. After consideration, our theoretical analysis reveals the existence of a wide class of estimators that are better than the standard one. Focusing on a subset of this class, we propose efficient shrinkage estimators for the kernel mean. Empirical evaluations on several applications clearly demonstrate that the proposed estimators outperform the standard kernel mean estimator.

1. Introduction

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

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

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

where $\mathcal{H}$ is an RKHS associated with a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Conditions ensuring that this expectation exists are given in Smola et al. (2007). Unfortunately, it is not practical to compute $\mu_{\mathbb{P}}$ directly because the distribution $\mathbb{P}$ is usually unknown. Instead, given an i.i.d sample $x_1, x_2, \ldots, x_n$ from $\mathbb{P}$, we can easily compute the empirical kernel mean by the average

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

The estimate $\widehat{\mu}_{\mathbb{P}}$ is the most commonly used estimate of the true kernel mean. Our primary interest here is to investigate whether one 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 Agnan, 2004; Smola et al., 2007). Representing the distribution as a mean function in the RKHS has several advantages: 1) the representation with appropriate choice of kernel k has been shown to preserve all information about the distribution (Fukumizu et al., 2004; Sriperumbudur et al., 2008; 2010); 2) 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 $f \in \mathcal{H}$; 3) no intermediate density estimation is required, e.g., when testing for homogeneity from finite samples. As a result, many algorithms have benefited from the kernel mean representation, namely, maximum mean discrepancy (MMD) (Gretton et al., 2007), kernel dependency measure (Gretton et al., 2005), kernel two-sample-test (Gretton et al., 2012), Hilbert space embedding of HMMs (Song et al., 2010), and kernel Bayes rule

(Fukumizu et al., 2011). Their performances rely directly on the quality of the empirical estimate $\widehat{\mu}_{\mathbb{P}}$.

However, it is of great importance, especially for our readers who are not familiar with kernel methods, to realize a more fundamental role of the kernel mean. It basically serves as a foundation to most kernel-based learning algorithms. For instance, nonlinear component analyses, such as kernel PCA, kernel FDA, and kernel CCA, rely heavily on mean functions and covariance operators in RKHS (Schölkopf et al., 1998). The kernel k-means algorithm performs clustering in feature space using mean functions as the representatives of the clusters (Dhillon et al., 2004). Moreover, it also serves as a basis in early development of algorithms for classification and anomaly detection (Shawe-Taylor and Cristianini, 2004, chap. 5). All of those employ (2) as the estimate of the true mean function. Thus, the fact that substantial improvement can be gained when estimating (1) may in fact raise a widespread suspicion on traditional way of learning with kernels.

We show in this work that the standard estimator (2) is, in a certain sense, not optimal, i.e., there exist better estimators (more below). In addition, we propose shrinkage estimators that outperform the standard one. At first glance, it was definitely counter-intuitive and surprising for us, and will undoubtedly also be for some of our readers, that the empirical kernel mean could be improved, and, given the simplicity of the proposed estimators, that this has remained unnoticed until now. One of the reasons may be that there is a common belief that the estimator $\hat{\mu}_{\mathbb{P}}$ already gives a good estimate of $\mu_{\mathbb{P}}$, and, as sample size goes to infinity, the estimation error disappears (Shawe-Taylor and Cristianini, 2004). As a result, no need is felt to improve the kernel mean estimation. However, given a finite sample, substantial improvement is in fact possible and several factors may come into play, as will be seen later in this work.

This work was partly inspired by Stein's seminal work in 1955, which showed that a maximum likelihood estimator (MLE), i.e., the standard empirical mean, for the mean of the multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ is inadmissible (Stein, 1955). That is, there exists an estimator that always achieves smaller total mean squared error regardless of the true θ, when the dimension is at least 3. Perhaps the best known estimator of such kind is James-Steins estimator (James and Stein, 1961). 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).

However, our work differs fundamentally from the Stein's seminal works and those along this line in two aspects. First, our setting is *non-parametric* in a sense that we do not assume any parametric form of the distribution, whereas most of traditional works focus on some specific distributions, e.g., Gaussian distribution. Second, our setting involves a *non-linear feature map* into a high-dimensional space, if not infinite. As a result, higher moments of the distribution may come into play. Thus, one cannot adopt Stein's setting straightforwardly. A direct generalization of James-Stein estimator to infinite-dimensional Hilbert space has already been considered (Berger and Wolpert, 1983; Mandelbaum and Shepp, 1987; Privault and Rveillac, 2008). In those works, $\boldsymbol{\theta}$ which is 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 our case, on the other hand, the samples are drawn from $\mathbb{P}$ and not from the Gaussian distribution whose mean is $\mu_{\mathbb{P}}$.

The contribution of this paper can be summarized as follows: First, we show that the standard kernel mean estimator can be improved by providing an alternative estimator that achieves smaller risk (§2). The theoretical analysis reveals the existence of a wide class of estimators that are better than the standard. To this end, we propose in §3 a *kernel mean shrinkage estimator* (KMSE), which is based on a novel motivation for regularization through the notion of shrinkage. Moreover, we propose an efficient leave-one-out cross-validation procedure to select the shrinkage parameter, which is novel in the context of kernel mean estimation. Lastly, we demonstrate the benefit of the proposed estimators in several applications (§4).

2. Motivation: Shrinkage Estimators

For an arbitrary distribution $\mathbb{P}$, denote by μ and $\widehat{\mu}$ the true kernel mean and its empirical estimate (2) from the i.i.d. sample $x_1, x_2, \ldots, x_n \sim \mathbb{P}$ (we remove the subscript for ease of notation). The most natural loss function considered in this work is $\ell(\mu, \widehat{\mu}) = \|\mu - \widehat{\mu}\|^2_{\mathcal{H}}$. An estimator $\widehat{\mu}$ is a mapping which is measurable w.r.t. the Borel σ-algebra of $\mathcal{H}$ and is evaluated by its risk function $\mathcal{R}(\mu, \widehat{\mu}) = \mathbb{E}_{\mathbb{P}}[\ell(\mu, \widehat{\mu})]$ where $\mathbb{E}_{\mathbb{P}}$ indicates expectation over the choice of i.i.d. sample of size n from $\mathbb{P}$.

Let us consider an alternative kernel mean estimator: $\widehat{\mu}_\alpha \triangleq \alpha f^* + (1-\alpha)\widehat{\mu}$ where $0 \leq \alpha < 1$ and $f^* \in \mathcal{H}$. It is essentially a shrinkage estimator that shrinks the standard estimator toward a function f^* by an amount specified by α. If $\alpha = 0$, $\widehat{\mu}_\alpha$ reduces to the standard estimator $\widehat{\mu}$. The following theorem asserts that the risk of shrinkage estimator $\widehat{\mu}_\alpha$ is smaller than that of standard estimator $\widehat{\mu}$ given an appropriate choice of α, regardless of the function f^* (more below).

Theorem 1. *For all distributions $\mathbb{P}$ and the kernel k, there exists $\alpha > 0$ for which $\mathcal{R}(\mu, \widehat{\mu}_\alpha) < \mathcal{R}(\mu, \widehat{\mu})$.*

Proof. The risk of the standard kernel mean estimator satisfies $\mathbb{E}\|\widehat{\mu} - \mu\|^2 = \frac{1}{n}\left(\mathbb{E}[k(x,x)] - \mathbb{E}[k(x,\tilde{x})]\right) =: \Delta$

where $\tilde{x}$ is an independent copy of x. Let us define the risk of the proposed shrinkage estimator by $\Delta_\alpha := \mathbb{E}\|\widehat{\mu}_\alpha - \mu\|^2$ where α is a non-negative shrinkage parameter. We can then write this in terms of the standard risk as $\Delta_\alpha = \Delta - 2\alpha\mathbb{E}\left\langle\widehat{\mu} - \mu, \widehat{\mu} - \mu + \mu - f^*\right\rangle + \alpha^2\mathbb{E}\|f^*\|^2 - 2\alpha^2\mathbb{E}[f^*(x)] + \alpha^2\mathbb{E}\|\widehat{\mu}\|^2$. It follows from the reproducing property of $\mathcal{H}$ that $\mathbb{E}[f^*(x)] = \langle f^*, \mu\rangle$. Moreover, using the fact that $\mathbb{E}\|\widehat{\mu}\|^2 = \mathbb{E}\|\widehat{\mu} - \mu + \mu\|^2 = \Delta + \mathbb{E}[k(x, \tilde{x})]$, we can simplify the shrinkage risk by $\Delta_\alpha = \alpha^2(\Delta + \|f^* - \mu\|^2) - 2\alpha\Delta + \Delta$. Thus, we have $\Delta_\alpha - \Delta = \alpha^2(\Delta + \|f^* - \mu\|^2) - 2\alpha\Delta$ which is non-positive where

$$\alpha \in \left[0, \frac{2\Delta}{\Delta + \|f^* - \mu\|^2}\right] \tag{3}$$

and minimized at $\alpha^* = \Delta/(\Delta + \|f^* - \mu\|^2)$. ■

As we can see in (3), there is a range of α for which a non-positive $\Delta_\alpha - \Delta$, i.e., $\mathcal{R}(\mu, \widehat{\mu}_\alpha) - \mathcal{R}(\mu, \widehat{\mu})$, is guaranteed. However, Theorem 1 relies on the important assumption that the true kernel mean of the distribution $\mathbb{P}$ is required to estimate α. In spite of this, the theorem has an important implication suggesting that the shrinkage estimator $\widehat{\mu}_\alpha$ can improve upon $\widehat{\mu}$ if α is chosen appropriately. Later, we will exploit this result in order to construct more practical estimators.

Remark 1. *The following observations follow immediately from Theorem 1:*

- *The shrinkage estimator always improves upon the standard one regardless of the direction of shrinkage, as specified by f^*. In other words, there exists a wide class of kernel mean estimators that are better than the standard one.*
- *The value of α also depends on the choice of f^*. The further f^* is from μ, the smaller 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.*
- *The improvement can be viewed as a bias-variance trade-off: the shrinkage estimator reduces variance substantially at the expense of a little bias.*

Remark 1 sheds light on how one can practically construct the shrinkage estimator: we can choose f^* arbitrarily as long as the parameter α is chosen appropriately. Moreover, further improvement can be gained by incorporating prior knowledge as to the location of $\mu_\mathbb{P}$, which can be straightforwardly integrated into the framework via f^* (Berger and Wolpert, 1983). Inspired by James-Stein estimator, we focus on $f^* = \mathbf{0}$. We will investigate the effect of different prior f^* in future works.

3. Kernel Mean Shrinkage Estimator

In this section we give a novel formulation of kernel mean estimator that allows us to estimate the shrinkage parameter efficiently. In the following, let $\phi : \mathcal{X} \to \mathcal{H}$ be a feature map associated with the kernel k and $\langle\cdot,\cdot\rangle$ be an inner product in the RKHS $\mathcal{H}$ such that $k(x, x') = \langle\phi(x), \phi(x')\rangle$. Unless stated otherwise, $\|\cdot\|$ denotes the RKHS norm. The kernel mean $\mu_\mathbb{P}$ and its empirical estimate $\widehat{\mu}_\mathbb{P}$ can be obtained as a minimizer of the loss functionals

$$\begin{aligned}\mathcal{E}(g) &\triangleq \mathbb{E}_{x\sim\mathbb{P}}\|\phi(x) - g\|^2,\\ \widehat{\mathcal{E}}(g) &\triangleq \frac{1}{n}\sum_{i=1}^n\|\phi(x_i) - g\|^2,\end{aligned}$$

respectively. We will call the estimator minimizing the loss functional $\widehat{\mathcal{E}}(g)$ a *kernel mean estimator (KME)*.

Note that the loss $\mathcal{E}(g)$ is different from the one considered in §2, i.e., $\ell(\mu, g) = \|\mu - g\|^2 = \|\mathbb{E}[\phi(x)] - g\|^2$. Nevertheless, we have $\ell(\mu, g) = \mathbb{E}_{xx'}k(x, x') - 2\mathbb{E}_x g(x) + \|g\|^2$. Since $\mathcal{E}(g) = \mathbb{E}_x k(x, x) - 2\mathbb{E}_x g(x) + \|g\|^2$, the loss $\ell(\mu, g)$ differs from $\mathcal{E}(g)$ only by $\mathbb{E}_x k(x, x) - \mathbb{E}_{xx'}k(x, x')$ which is not a function of g. We introduce the new form here because it will give a more tractable cross-validation computation (§3.1). In spite of this, the resulting estimators are always evaluated w.r.t. the loss in §2 (cf. §4.1).

From the formulation above, it is natural to ask if minimizing the regularized version of $\widehat{\mathcal{E}}(g)$ will give better estimator. On the one hand, one can argue that, unlike in the classical risk minimization, we do not really need a regularizer here. The standard estimator (2) is known to be, in a certain sense, optimal and can be estimated reliably (Shawe-Taylor and Cristianini, 2004, prop. 5.2). Moreover, the original formulation of $\widehat{\mathcal{E}}(g)$ is a well-posed problem. On the other hand, since regularization may be viewed as shrinking the solution toward zero, it can actually improve the kernel mean estimation, as suggested by Theorem 1 (cf. discussions at the end of §2).

Consequently, we minimize a modified loss functional

$$\begin{aligned}\widehat{\mathcal{E}}_\lambda(g) &\triangleq \widehat{\mathcal{E}}(g) + \lambda\Omega(\|g\|)\\ &= \frac{1}{n}\sum_{i=1}^n\|\phi(x_i) - g\|^2 + \lambda\Omega(\|g\|),\end{aligned} \tag{4}$$

where $\Omega(\cdot)$ denotes a monotonically-increasing regularization functional and λ is a non-negative regularization parameter.[1] In what follows, we refer to the shrinkage estimator $\widehat{\mu}_\lambda$ minimizing $\widehat{\mathcal{E}}_\lambda(g)$ as a *kernel mean shrinkage estimator (KMSE)*.

[1] The parameters α and λ play similar role as a shrinkage parameter. They specify an amount by which the standard estimator $\widehat{\mu}$ is shrunk toward $f^* = 0$. Thus, the term shrinkage parameter and regularization parameter will be used interchangeably.

It follows from the representer theorem that g lies in a subspace spanned by the data, i.e., $g = \sum_{j=1}^n \beta_j \phi(x_j)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. By considering $\Omega(\|g\|) = \|g\|^2$, we can rewrite (4) as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \left\| \phi(x_i) - \sum_{j=1}^n \beta_j \phi(x_j) \right\|^2 + \lambda \left\| \sum_{j=1}^n \beta_j \phi(x_j) \right\|^2 \\ &= \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{K} \mathbf{1}_n + \lambda \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} + c, \end{aligned} \tag{5}$$

where c is a constant term, $\mathbf{K}$ is an $n \times n$ Gram matrix such that $\mathbf{K}_{ij} = k(x_i, x_j)$, and $\mathbf{1}_n = [1/n, 1/n, \ldots, 1/n]^\top$. Taking a derivative of (5) w.r.t. $\boldsymbol{\beta}$ and setting it to zero yield $\boldsymbol{\beta} = (1/(1+\lambda))\mathbf{1}_n$. By setting $\alpha = \lambda/(1+\lambda)$ the shrinkage estimate can be written as $\widehat{\mu}_\lambda = (1-\alpha)\widehat{\mu}$. Since $0 < \alpha < 1$, the estimator $\widehat{\mu}_\lambda$ corresponds to a shrinkage estimator discussed in §2 when $f^* = \mathbf{0}$. We call this estimator a *simple kernel mean shrinkage estimator (S-KMSE)*.

Using the expansion $g = \sum_{j=1}^n \beta_j \phi(x_j)$, we may consider when the regularization functional is written in term of $\boldsymbol{\beta}$, e.g., $\boldsymbol{\beta}^\top \boldsymbol{\beta}$. This leads to a particularly interesting kernel mean estimator. In this case, the optimal weight vector is given by $\boldsymbol{\beta} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n$ and the shrinkage estimate can be written accordingly as $\widehat{\mu}_\lambda = \sum_{j=1}^n \beta_j \phi(x_j) = \Phi^\top (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{K} \mathbf{1}_n$ where $\Phi = [\phi(x_1), \phi(x_2), \ldots, \phi(x_n)]^\top$. Unlike the S-KMSE, this estimator shrinks the usual estimate differently in each coordinate (cf. Theorem 2). Hence, we will call it a *flexible kernel mean shrinkage estimator (F-KMSE)*.

The following theorem characterizes the F-KMSE as a shrinkage estimator.

Theorem 2. *The F-KMSE can be written as* $\widehat{\mu}_\lambda = \sum_{i=1}^n \frac{\gamma_i}{\gamma_i + \lambda} \langle \widehat{\mu}, \mathbf{v}_i \rangle \mathbf{v}_i$ *where* $\{\gamma_i, \mathbf{v}_i\}$ *are eigenvalue and eigenvector pairs of the empirical covariance operator* $\widehat{\mathbf{C}}_{xx}$ *in* $\mathcal{H}$.

In words, the effect of F-KMSE is to reduce high frequency components of the expansion of $\widehat{\mu}_\lambda$, by expanding this in terms of the kernel PCA basis and shrinking the coefficients of the high order eigenfunctions, e.g., see Rasmussen and Williams (2006, sec. 4.3). Note that the covariance operator $\widehat{\mathbf{C}}_{xx}$ itself does not depend on λ.

As we can see, the solution to the regularized version is indeed of the form of shrinkage estimators when $f^* = \mathbf{0}$. That is, both S-KMSE and F-KMSE shrink the standard kernel mean estimate towards zero. The difference is that the S-KMSE shrinks equally in all coordinate, whereas the F-KMSE also constraints the amount of shrinkage by the information contained in each coordinate.

Moreover, the squared RKHS norm $\|\cdot\|^2$ can be decomposed as a sum of squared loss weighted by the eigenvalues γ_i (cf. Mandelbaum and Shepp (1987, appendix)). By the same reasoning as Stein's result in finite-dimensional case, one would suspect that an improvement of shrinkage estimators in $\mathcal{H}$ should also depend on how fast the eigenvalues of k decay. That is, one would expect greater improvement if the values of γ_i decay very slowly. For example, the Gaussian RBF kernel with larger bandwidth gives smaller improvement when compared to one with smaller bandwidth. Similarly, we should expect to see more improvement when applying a Laplacian kernel than when using a Gaussian RBF kernel.

In some applications of kernel mean embedding, one may want to interpret the weight $\boldsymbol{\beta}$ as a probability vector (Nishiyama et al., 2012). However, the weight vector $\boldsymbol{\beta}$ output by our estimators is in general not normalized. In fact, all elements will be smaller than $1/n$ as a result of shrinkage. However, one may impose a constraint that $\boldsymbol{\beta}$ must sum to one and resort to a quadratic programming (Song et al., 2008). Unfortunately, this approach has undesirable effect of sparsity which is unlikely to improve upon the standard estimator. Post-normalizing the weights often deteriorates the estimation performance.

To the best of our knowledge, no previous attempt has been made to improve the kernel mean estimation. However, we discuss some closely related works here. For example, instead of the loss functional $\widehat{\mathcal{E}}(g)$, Kim and Scott (2012) consider a robust loss function such as the Huber's loss to reduce the effect of outliers. The authors consider kernel density estimators, which differ fundamentally from kernel mean estimators. They need to reduce the kernel bandwidth with increasing sample size for the estimators to be consistent. Regularized version of MMD was adopted by Danafar et al. (2013) in the context of kernel-based hypothesis testing. The resulting formulation resembles our S-KMSE. Furthermore, the F-KMSE is of a similar form as the conditional mean embedding used in Grünewälder et al. (2012), which can be viewed more generally as a regression problem in RKHS with smooth operators (Grünewälder et al., 2013).

3.1. Choosing Shrinkage Parameter λ

As discussed in §2, the amount of shrinkage plays an important role in our estimators. In this work we propose to select the shrinkage parameter λ by an automatic leave-one-out cross-validation.

For a given shrinkage parameter λ, let us consider the observation x_i as being a new observation by omitting it from the dataset. Denote by $\widehat{\mu}_\lambda^{(-i)} = \sum_{j \neq i} \beta_j^{(-i)} \phi(x_j)$ the kernel mean estimated from the remaining data, using the value λ as a shrinkage parameter, so that $\boldsymbol{\beta}^{(-i)}$ is the minimizer of $\widehat{\mathcal{E}}_\lambda^{(-i)}(g)$. We will measure the quality of $\widehat{\mu}_\lambda^{(-i)}$ by how well it approximates $\phi(x_i)$. The overall quality of the

estimate is quantified by the cross-validation score

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\|\phi(x_i) - \widehat{\mu}_\lambda^{(-i)}\right\|_{\mathcal{H}}^2. \qquad (6)$$

By simple algebra, it is not difficult to show that the optimal shrinkage parameter of S-KMSE can be calculated analytically, as stated by the following theorem.

Theorem 3. *Let $\rho \triangleq \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,x_j)$ and $\varrho \triangleq \frac{1}{n}\sum_{i=1}^n k(x_i,x_i)$. The shrinkage parameter $\lambda_* = (\varrho - \rho)/((n-1)\rho + \varrho/n - \varrho)$ of the S-KMSE is the minimizer of $LOOCV(\lambda)$.*

On the other hand, finding the optimal λ for the F-KMSE is relatively more involved. Evaluating the score (6) naïvely requires one to solve for $\widehat{\mu}_\lambda^{(-i)}$ explicitly for every i. Fortunately, we can simplify the score such that it can be evaluated efficiently, as stated in the following theorem.

Theorem 4. *The LOOCV score of F-KMSE satisfies $LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_i)^\top\mathbf{C}_\lambda(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_i)$ where $\boldsymbol{\beta}$ is the weight vector calculated from the full dataset with the shrinkage parameter λ and $\mathbf{C}_\lambda = (\mathbf{K} - \frac{1}{n}\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{K})^{-1}\mathbf{K}(\mathbf{K} - \frac{1}{n}\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{K})^{-1}$.*

Proof of Thorem 4. For fixed λ and i, let $\widehat{\mu}_\lambda^{(-i)}$ be the leave-one-out kernel mean estimate of F-KMSE and let $\mathbf{A} \triangleq (\mathbf{K}+\lambda\mathbf{I})^{-1}$. Then, we can write an expression for the deleted residual as $\Delta_\lambda^{(-i)} := \widehat{\mu}_\lambda^{(-i)} - \phi(x_i) = \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^n\sum_{l=1}^n \mathbf{A}_{jl}\langle\phi(x_l), \widehat{\mu}_\lambda^{(-i)} - \phi(x_i)\rangle\phi(x_j)$. Since $\Delta_\lambda^{(-i)}$ lies in a subspace spanned by the sample $\phi(x_1),\ldots,\phi(x_n)$, we have $\Delta_\lambda^{(-i)} = \sum_{k=1}^n \xi_k\phi(x_k)$ for some $\boldsymbol{\xi} \in \mathbb{R}^n$. Substituting $\Delta_\lambda^{(-i)}$ back yields $\sum_{k=1}^n\xi_k\phi(x_k) = \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^n\{\mathbf{AK}\boldsymbol{\xi}\}_j\phi(x_j)$. By taking the inner product on both sides w.r.t. the sample $\phi(x_1),\ldots,\phi(x_n)$ and solving for $\boldsymbol{\xi}$, we have $\boldsymbol{\xi} = (\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i})$ where $\mathbf{K}_{\cdot i}$ is the ith column of $\mathbf{K}$. Consequently, the leave-one-out score of the sample x_i can be computed by $\|\Delta_\lambda^{(-i)}\|^2 = \boldsymbol{\xi}^\top\mathbf{K}\boldsymbol{\xi} = (\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i})^\top(\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}\mathbf{K}(\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i}) = (\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i})^\top\mathbf{C}_\lambda(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i})$. Averaging $\|\Delta_\lambda^{(-i)}\|^2$ over all samples gives $LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^n\|\Delta_\lambda^{(-i)}\|^2 = \frac{1}{n}\sum_{i=1}^n(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i})^\top\mathbf{C}_\lambda(\boldsymbol{\beta}^\top\mathbf{K} - \mathbf{K}_{\cdot i})$, as required. ■

It is interesting to see that the leave-one-out cross-validation score in Theorem 4 depends only on the non-leave-one-out solution $\boldsymbol{\beta}_\lambda$, which can be obtained as a by-product of the algorithm.

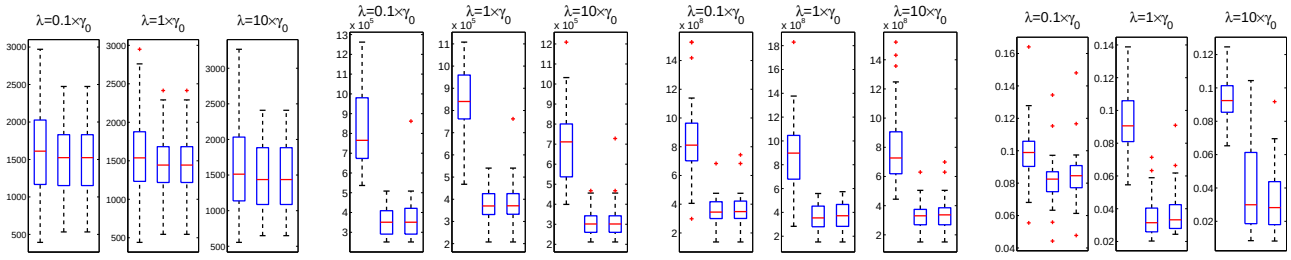
Computational complexity The S-KMSE requires $\mathcal{O}(n^2)$ operations to select shrinkage parameter. For the F-KMSE, there are two steps in cross-validation. First, we need to compute $(\mathbf{K}+\lambda\mathbf{I})^{-1}$ repeatedly for different values of λ. Assume that we know the eigendecomposition $\mathbf{K} = \mathbf{UDU}^\top$ where $\mathbf{D}$ is diagonal with $d_{ii} \geq 0$ and $\mathbf{UU}^\top = \mathbf{I}$. It follows that $(\mathbf{K}+\lambda\mathbf{I})^{-1} = \mathbf{U}(\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{U}^\top$. Consequently, solving for $\boldsymbol{\beta}_\lambda$ takes $\mathcal{O}(n^2)$ operations. Since eigendecomposition requires $\mathcal{O}(n^3)$ operations, finding $\boldsymbol{\beta}_\lambda$ for many λ's is essentially free. A low-rank approximation can also be adopted to reduce the computational cost further.

Second, we need to compute the cross-validation score (6). As shown in Theorem 4, we can compute it using only $\boldsymbol{\beta}_\lambda$ obtained from the previous step. The calculation of $\mathbf{C}_\lambda$ can be simplified further via the eigendecomposition of $\mathbf{K}$ as $\mathbf{C}_\lambda = \mathbf{U}(\mathbf{D} - \frac{1}{n}\mathbf{D}(\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D})^{-1}\mathbf{D}(\mathbf{D} - \frac{1}{n}\mathbf{D}(\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{D})^{-1}\mathbf{U}^\top$. Since it only involves the inverse of diagonal matrices, the inversion can be evaluated in $\mathcal{O}(n)$ operations. The overall computational complexity of the cross-validation requires only $\mathcal{O}(n^2)$ operations, as opposed to the naïve approach that requires $\mathcal{O}(n^4)$ operations. When performed as a by-product of the algorithm, the computational cost of cross-validation procedure becomes negligible as the dataset becomes larger. In practice, we use the `fminsearch` and `fminbnd` routines of the MATLAB optimization toolbox to find the best shrinkage parameter.

3.2. Covariance Operators

The covariance operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ can be viewed as a mean function in a product space $\mathcal{H}_X \otimes \mathcal{H}_Y$. Hence, we can also construct a shrinkage estimator of covariance operator in RKHS. Let $(\mathcal{H}_X, k_X)$ and $(\mathcal{H}_Y, k_Y)$ be the RKHS of functions on measurable space $\mathcal{X}$ and $\mathcal{Y}$, respectively, with p.d. kernel k_X and k_Y (with feature map ϕ and φ). We will consider a random vector $(X,Y) : \Omega \to \mathcal{X}\times\mathcal{Y}$ with distribution $\mathbb{P}_{XY}$, with $\mathbb{P}_X$ and $\mathbb{P}_Y$ as marginal distributions. Under some conditions, 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 $g \in \mathcal{H}_Y$ (Fukumizu et al., 2004). If X equals Y, we get the self-adjoint operator Σ_{XX} called the covariance operator.

Given an i.i.d sample from $\mathbb{P}_{XY}$ written as $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, we can write the empirical cross-covariance operator as $\widehat{\Sigma}_{YX} := \frac{1}{n}\sum_{i=1}^n\phi(x_i)\otimes\varphi(y_i) - \widehat{\mu}_X\otimes\widehat{\mu}_Y$ where $\widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^n\phi(x_i)$ and $\widehat{\mu}_Y = \frac{1}{n}\sum_{i=1}^n\varphi(y_i)$. Let $\widetilde{\phi}$ and $\widetilde{\varphi}$ be the centered feature maps of ϕ and φ, respectively. Then, it can be rewritten as $\widehat{\Sigma}_{YX} := \frac{1}{n}\sum_{i=1}^n\widetilde{\phi}(x_i)\otimes\widetilde{\varphi}(y_i) \in \mathcal{H}_X\otimes\mathcal{H}_Y$. It follows from the inner product property in product space that $\langle\widetilde{\phi}(x)\otimes\widetilde{\varphi}(y), \widetilde{\phi}(x')\otimes\widetilde{\varphi}(y')\rangle_{\mathcal{H}_X\otimes\mathcal{H}_Y} = \langle\widetilde{\phi}(x),\widetilde{\phi}(x')\rangle_{\mathcal{H}_X}\langle\widetilde{\varphi}(y),\widetilde{\varphi}(y')\rangle_{\mathcal{H}_Y} = \widetilde{k}_X(x,x')\widetilde{k}_Y(y,y')$.



(a) LIN (b) POLY2 (c) POLY3 (d) RBF

Figure 1. The average loss of KME (left), S-KMSE (middle), and F-KMSE (right) estimators with different values of shrinkage parameter. Inside boxes correspond to estimators. We repeat the experiments over 30 different distributions with $n = 10$ and $d = 30$.

Then, we can obtain the shrinkage estimators for the covariance operator by plugging the kernel $k((x, y), (x', y')) = \tilde{k}_X(x, x')\tilde{k}_Y(y, y')$ in our KMSEs. We will call this estimator a *covariance-operator shrinkage estimator (COSE)*. The same trick can be easily generalized to tensors of higher order, which have been previously used, for example, in Song et al. (2011).

4. Experiments

We focus on the comparison between our shrinkage estimators and the standard estimator of the kernel mean using both synthetic datasets and real-world datasets.

4.1. Synthetic Data

Given the true data-generating distribution $\mathbb{P}$, we evaluate different estimators using the loss function $\ell(\boldsymbol{\beta}) \triangleq \|\sum_{i=1}^{n} \beta_i k(x_i, \cdot) - \mathbb{E}_{\mathbb{P}}[k(x, \cdot)]\|^2_{\mathcal{H}}$ where $\boldsymbol{\beta}$ is the weight vector associated with different estimators. To allow for an exact calculation of $\ell(\boldsymbol{\beta})$, we consider when $\mathbb{P}$ is a mixture-of-Gaussians distribution and k is the following kernel function: 1) linear kernel $k(x, x') = x^\top x'$; 2) polynomial degree-2 kernel $k(x, x') = (x^\top x' + 1)^2$; 3) polynomial degree-3 kernel $k(x, x') = (x^\top x' + 1)^3$; and 4) Gaussian RBF kernel $k(x, x') = \exp\left(-\|x - x'\|^2/2\sigma^2\right)$. We will refer to them as LIN, POLY2, POLY3, and RBF, respectively.

Experimental protocol. Data are generated from a d-dimensional mixture of Gaussians:

$$x \sim \sum_{i=1}^{4} \pi_i \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_i) + \varepsilon, \quad \theta_{ij} \sim \mathcal{U}(-10, 10),$$
$$\Sigma_i \sim \mathcal{W}(2 \times \mathbf{I}_d, 7), \qquad \varepsilon \sim \mathcal{N}(0, 0.2 \times \mathbf{I}_d),$$

where $\mathcal{U}(a, b)$ and $\mathcal{W}(\Sigma_0, df)$ represent the uniform distribution and Wishart distribution, respectively. We set $\boldsymbol{\pi} = [0.05, 0.3, 0.4, 0.25]$. The choice of parameters here is quite arbitrary; we have experimented using various parameter settings and the results are similar to those presented here. For the Gaussian RBF kernel, we set the bandwidth parameter to square-root of the median Euclidean distance between samples in the dataset (i.e., $\sigma^2 = \text{median}\left\{\|x_i - x_j\|^2\right\}$ throughout).

Figure 1 shows the average loss of different estimators using different kernels as we increase the value of shrinkage parameter λ. Here we scale the shrinkage parameter by the minimum non-zero eigenvalue γ_0 of kernel matrix $\mathbf{K}$. In general, we find S-KMSE and F-KMSE tend to outperform KME. However, as λ becomes large, there are some cases where shrinkage deteriorates the estimation performance, e.g., see LIN kernel and some outliers in the figures. This suggests that it is very important to choose the parameter λ appropriately (cf. the discussion in §2).

Similarly, Figure 2 depicts the average loss as we vary the sample size and dimension of the data. In this case, the shrinkage parameter is chosen by the proposed leave-one-out cross-validation score. As we can see, both S-KMSE

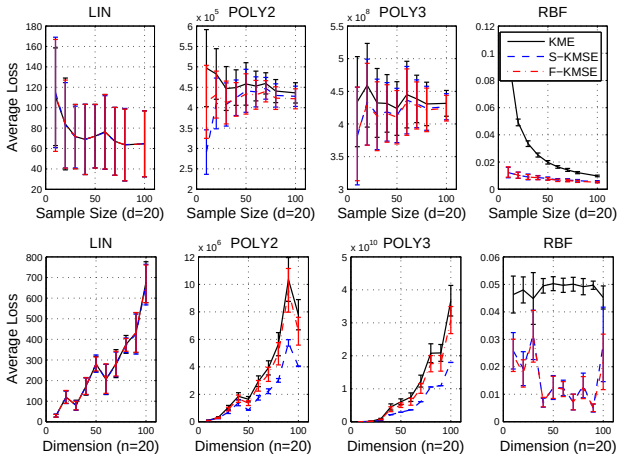


Figure 2. The average loss over 30 different distributions of KME, S-KMSE, and F-KMSE with varying sample size (n) and dimension (d). The shrinkage parameter λ is chosen by LOOCV.

Table 1. Average negative log-likelihood of the model Q on test points over 10 randomizations. The boldface represents the result whose difference from the baseline, i.e., KME, is statistically significant.

Dataset		LIN			POLY2			POLY3			RBF		
		KME	S-KMSE	F-KMSE	KME	S-KMSE	F-KMSE	KME	S-KMSE	F-KMSE	KME	S-KMSE	F-KMSE
1.	ionosphere	33.2440	33.0325	33.1436	53.1266	53.7067	50.8695	51.6800	49.9149	**47.4461**	40.8961	**40.5578**	**39.6804**
2.	sonar	72.6630	72.8770	72.5015	120.3454	**108.8246**	**109.9980**	102.4499	**90.3920**	91.1547	71.3048	70.5721	**70.5830**
3.	australian	18.3703	**18.3341**	18.3719	18.5928	18.6028	18.4987	41.1563	**34.4303**	**34.5460**	17.5138	17.5637	17.4026
4.	specft	56.6138	55.7374	**55.8667**	67.3901	65.9662	65.2056	63.9273	63.5571	**62.1480**	57.5569	**56.1386**	**55.5808**
5.	wdbc	30.9778	30.9266	**30.4400**	93.0541	91.5803	**87.5265**	58.8235	54.1237	**50.3911**	30.8227	**30.5968**	**30.2646**
6.	wine	15.9225	15.8850	16.0431	24.2841	24.1325	**23.5163**	35.2069	32.9465	32.4702	17.1523	16.9177	**16.6312**
7.	satimage*	19.6353	19.8721	19.7943	149.5986	143.2277	146.0648	52.7973	57.2482	45.8946	20.3306	20.5020	**20.2226**
8.	segment	22.9131	22.8219	**22.0696**	61.2712	59.4387	**54.8621**	38.7226	38.6226	38.4217	17.6801	**16.4149**	**15.6814**
9.	vehicle	16.4145	16.2888	16.3210	83.1597	**79.7248**	**79.6679**	70.4340	63.4322	48.0177	15.9256	15.8331	**15.6516**
10.	svmguide2	27.1514	**27.0644**	27.1144	30.3065	30.2290	29.9875	37.0427	36.7854	35.8157	27.3930	**27.2517**	**27.1815**
11.	vowel	12.4227	12.4219	12.4264	32.1389	**28.0474**	**29.3492**	25.8728	**24.0684**	**23.9747**	12.3976	12.3823	12.3677
12.	housing	15.5249	15.1618	15.3176	39.9582	37.1360	32.1028	50.8481	49.0884	35.1366	14.5576	14.3810	**13.9379**
13.	bodyfat	17.6426	**17.0419**	17.2152	44.3295	43.7959	**42.3331**	27.4339	25.6530	24.7955	16.2725	**15.9170**	**15.8665**
14.	abalone*	4.3348	4.3274	4.3187	14.9166	14.4041	11.4431	20.6071	23.2487	23.6291	4.6928	4.6056	4.6017
15.	glass	10.4078	10.4451	**10.4067**	33.3480	31.6110	30.5075	45.0801	34.9608	**25.5677**	8.6167	8.4992	**8.2469**

and F-KMSE outperform the standard KME. The S-KMSE performs slightly better than the F-KMSE. Moreover, the improvement is more substantial in the "large d, small n" paradigm. In the worst cases, the S-KMSE and F-KMSE perform as well as the KME.

Lastly, it is instructive to note that the improvement varies with the choice of kernel k. Briefly, the choice of kernel reflects the dimensionality of feature space $\mathcal{H}$. One would expect more improvement in high-dimensional space, e.g., RBF kernel, than the low-dimensional, e.g., linear kernel (cf. discussions at the end of §3). This phenomenon can be observed in both Figure 1 and 2.

4.2. Real Data

We consider three benchmark applications: density estimation via kernel mean matching (Song et al., 2008), kernel PCA using shrinkage mean and covariance operator (Schölkopf et al., 1998), and discriminative learning on distributions (Muandet and Schölkopf, 2013; Muandet et al., 2012). For the first two tasks we employ 15 datasets from the UCI repositories. We use only real-valued features, each of which is normalized to have zero mean and unit variance.

Density estimation. We perform density estimation via kernel mean matching (Song et al., 2008). That is, 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 minimizing $\|\widehat{\mu} - \mu_Q\|_{\mathcal{H}}^2$ s.t. $\sum_{j=1}^{m} \pi_j = 1$. The kernel mean $\widehat{\mu}$ is obtained from the 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. That is, we replace $\hat{\mu}$ with a version obtained via shrinkage. 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. We use 30% of the dataset as a test set. We set $m = 10$ for each dataset. The model is initialized by running 50 random initializations using the k-means algorithm and returning the best. We repeat the experiments 10 times and perform the paired sign test on the results at the 5% significance level.[2]

The average negative log-likelihood of the model Q, optimized via different estimators, is reported in Table 1. Clearly, both S-KMSE and F-KMSE consistently achieve smaller negative log-likelihood when compared to KME. There are however few cases in which KME outperforms the proposed estimators, especially when the dataset is relatively large, e.g., `satimage` and `abalone`. We suspect that in those cases the standard KME already provides an accurate estimate of the kernel mean. To get a better estimate, more effort is required to optimize for the shrinkage parameter. Moreover, the improvement across different kernels is consistent with results on the synthetic datasets.

Kernel PCA. In this experiment, we perform the KPCA using different estimates of the mean and covariance operators. We compare the reconstruction error $\mathcal{E}_{proj}(z) = \|\phi(z) - \mathbf{P}\phi(z)\|^2$ on test samples where $\mathbf{P}$ is the projection constructed from the first 20 principal components. We use a Gaussian RBF kernel for all datasets. We compare 5 different scenarios: 1) standard KPCA; 2) shrinkage centering with S-KMSE; 3) shrinkage centering with F-KMSE; 4) KPCA with S-COSE; and 5) KPCA with F-COSE. To perform KPCA on shrinkage covariance operator, we solve the generalized eigenvalue problem $\mathbf{K}^c\mathbf{B}\mathbf{K}^c\mathbf{V} = \mathbf{K}^c\mathbf{V}\mathbf{D}$ where $\mathbf{B} = \mathrm{diag}(\boldsymbol{\beta})$ and $\mathbf{K}^c$ is the centered Gram matrix. The weight vector $\boldsymbol{\beta}$ is obtained from shrinkage estimators using the kernel matrix $\mathbf{K}^c \circ \mathbf{K}^c$ where $\circ$ denotes the Hadamard product. We use 30% of the dataset as a test set.

[2]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 S-KMSE and F-KMSE against KME.

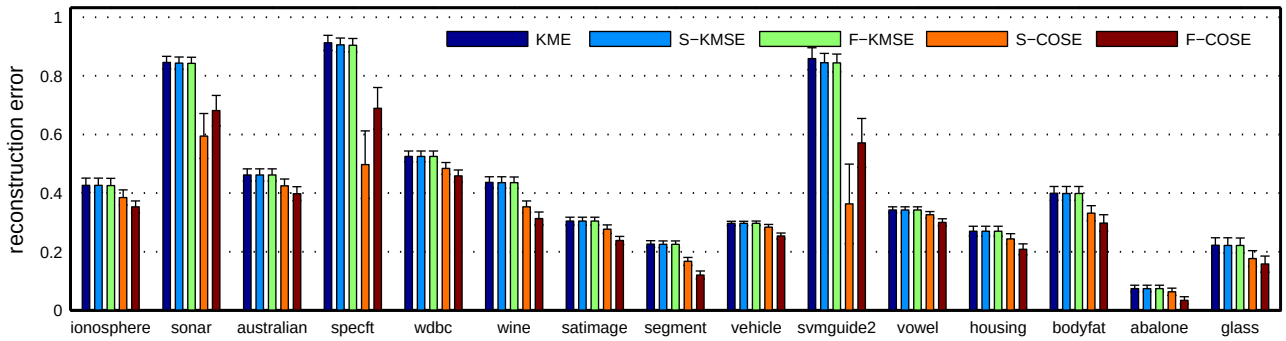


Figure 3. The average reconstruction error of KPCA on hold-out test samples over 10 repetitions. The KME represents the standard approach, whereas S-KMSE and F-KMSE use shrinkage means to perform centering. The S-COSE and F-COSE directly use the shrinkage estimate of the covariance operator.

Figure 3 illustrates the results of KPCA. Clearly, the S-COSE and F-COSE consistently outperforms all other estimators. Although we observe an improvement of S-KMSE and F-KMSE over KME, it is very small compared to that of S-COSE and F-COSE. This makes sense intuitively, since changing the mean point or shifting data does not change the covariance structure considerably, so it will not significantly affect the reconstruction error.

Discriminative learning on distributions. 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 $\widehat{\mathbb{P}}_i := \frac{1}{n}\sum_{k=1}^n \delta_{x_k^i}$ and $x_k^i \sim \mathbb{P}_i$, the linear kernel between two distributions is approximated by $\langle \widehat{\mu}_{\mathbb{P}_i}, \widehat{\mu}_{\mathbb{P}_j} \rangle = \langle \sum_{k=1}^n \beta_k^i \phi(x_k^i), \sum_{l=1}^n \beta_l^j \phi(x_l^j) \rangle = \sum_{k,l=1}^n \beta_k^i \beta_l^j k(x_k^i, x_l^j)$. 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., $\kappa(\mathbb{P}_i, \mathbb{P}_j) = \exp(\|\widehat{\mu}_{\mathbb{P}_i} - \widehat{\mu}_{\mathbb{P}_j}\|^2_{\mathcal{H}}/2\sigma^2)$. Our goal in this experiment is to investigate if the shrinkage estimate of the kernel mean improves the performance of the 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. For our unsupervised problem, we repeat the experiments using several parameter settings and report the best results.

Table 2 reports the classification accuracy of SMM and the area under ROC curve (AUC) of OCSMM using different kernel mean estimators. Both shrinkage estimators consistently lead to better performance on both SMM and OCSMM when compared to KME.

Table 2. The classification accuracy of SMM and the area under ROC curve (AUC) of OCSMM using different kernel mean estimators to construct the kernel on distributions.

Estimator	**Linear**		**Non-linear**	
	SMM	OCSMM	SMM	OCSMM
KME	0.5432	0.6955	0.6017	0.9085
S-KMSE	0.5521	0.6970	0.6303	0.9105
F-KMSE	0.5610	0.6970	0.6522	0.9095

To summarize, we find sufficient evidence to conclude that both S-KMSE and F-KMSE outperforms the standard KME. The performance of S-KMSE and F-KMSE is very competitive. The difference depends on the dataset and the kernel function.

5. Conclusions

To conclude, we show that the commonly used kernel mean estimator can be improved. Our theoretical result suggests that there exists a wide class of kernel mean estimators that are better than the standard one. To demonstrate this, we focus on two efficient shrinkage estimators, namely, simple and flexible kernel mean shrinkage estimators. Empirical study clearly shows that the proposed estimators outperform the standard one in various scenarios. Most importantly, the shrinkage estimates not only provide more accurate estimation, but also lead to superior performance on real-world applications.

Acknowledgments

The authors wish to thank David Hogg and Ross Fedely for reading the first draft and anonymous reviewers who gave valuable suggestion that has helped to improve the manuscript.

References

J. Berger and R. Wolpert. Estimating the mean function of a gaussian process and the stein effect. *Journal of Multivariate Analysis*, 13(3):401–424, 1983.

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

A. Berlinet and T. C. Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

S. Danafar, P. M. V. Rancoita, T. Glasmachers, K. Whittingstall, and J. Schmidhuber. Testing hypotheses by regularized maximum mean discrepancy. 2013.

I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 551–556, New York, NY, USA, 2004.

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

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

A. Gretton, R. Herbrich, A. Smola, B. Schölkopf, and A. Hyvärinen. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

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

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

S. Grünewälder, G. Lever, A. Gretton, L. Baldassarre, S. Patterson, and M. Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

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

W. James and J. 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.

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

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

K. Muandet and B. 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.

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

Y. Nishiyama, A. Boularias, A. Gretton, and K. Fukumizu. Hilbert space embeddings of POMDPs. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 644–653, 2012.

N. Privault and A. Rveillac. Stein estimation for the drift of gaussian processes using the malliavin calculus. *Annals of Statistics*, 36(5):2531–2550, 2008.

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

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, July 1998.

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

A. Smola, A. Gretton, L. Song, and B. 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.

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

L. Song, B. Boots, S. M. Siddiqi, G. J. Gordon, and A. J. Smola. Hilbert space embeddings of hidden Markov models. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

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

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *COLT*, 2008.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 99:1517–1561, 2010.

C. 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.