Supplementary Material to Kernel Mean Estimation and Stein Effect

Krikamol Muandet*[1], Kenji Fukumizu†[2], Bharath Sriperumbudur‡[3], Arthur Gretton§[4], and Bernhard Schölkopf¶[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]The Institute of Statistical Mathematics, Tokyo, Japan
[3]Statistical Laboratory, University of Cambridge, Cambridge, United Kingdom
[4]Gatsby Unit, University College London, London, United Kingdom

1 James-Stein Estimator

Stein's result has transformed common belief in statistical world that the maximum likelihood estimator, which is in common use for more than a century, is optimal. Charles Stein showed in 1955 that it is possible to uniformly improve the maximum likelihood estimator (MLE) for the Gaussian model in terms of total squared error risk when several parameters are estimated simultaneously from independent normal observations (Stein 1955). James and Stein later proposed a particularly simple estimator which dominates the usual MLE, given that there are more than two parameters (James and Stein 1961).

The following proposition gives a general form of the James-Stein estimator.

Proposition 1. *Assuming $X \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ with $dim(X) \geq 3$, the estimator $\delta(X) = X$ for $\boldsymbol{\theta}$ is inadmissible under the squared loss function and is dominated by the following estimator*

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

where d is the dimension of X.

Although the original works on James-Stein estimator were entirely written from the frequentist point of view, it was shown later that James-Stein estimator can be understood as an empirical Bayes estimator (Efron and Morris 1973a). This is a treatment of

*krikamol@tuebingen.mpg.de
†fukumizu@ism.ac.jp
‡bs493@statslab.cam.ac.uk
§arthur.gretton@gmail.com
¶bs@tuebingen.mpg.de

James-Stein estimator from the Bayesian point of view. There have been a considerable number of works in this direction, e.g., (Efron and Morris 1972; 1973b; 1975) and later by Berger (1975), Bock (1975), Hudson (1978). Whether the same Bayesian interpretation is possible in an infinite-dimensional space such as the RKHS is still an open problem.

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

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

where $\lambda \in [0,1]$, $\hat{\theta}_{ML}$ 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 $\tilde{\theta} = \mathbf{0}$. That is, it shrinks the usual estimator toward zero.

2 Proof of Theorem 1

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 standard kernel mean estimator satisfies

$$\mathbb{E}\|\widehat{\mu} - \mu\|^2 = \frac{1}{n}\left(\mathbb{E}[k(x,x)] - \mathbb{E}[k(x,\tilde{x})]\right) =: \Delta.$$

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. Then we can write it in term of the standard risk as follows:

$$\begin{aligned}
\Delta_\alpha &= \mathbb{E}\|(1-\alpha)\widehat{\mu} + \alpha f^* - \mu\|^2 \\
&= \mathbb{E}\|(\widehat{\mu} - \mu) + \alpha(f^* - \widehat{\mu})\|^2 \\
&= \Delta - 2\alpha\mathbb{E}\left\langle \widehat{\mu} - \mu, \widehat{\mu} - f^* \right\rangle + \alpha^2\mathbb{E}\|f^* - \widehat{\mu}\|^2 \\
&= \Delta - 2\alpha\mathbb{E}\left\langle \widehat{\mu} - \mu, \widehat{\mu} - \mu + \mu - f^* \right\rangle + \alpha^2\mathbb{E}\|f^*\|^2 - 2\alpha^2\mathbb{E}[f^*(x)] + \alpha^2\mathbb{E}\|\widehat{\mu}\|^2.
\end{aligned}$$

It follows from the reproducing property of $\mathcal{H}$ that $\mathbb{E}[f^*(x)] = \langle f^*, \mu\rangle$. 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 risk of shrinkage estimator by

$$\begin{aligned}
\Delta_\alpha &= \Delta - 2\alpha\Delta + \alpha^2\mathbb{E}\|f^*\|^2 - 2\alpha^2\left\langle f^*, \mu\right\rangle + \alpha^2\left(\Delta + \mathbb{E}[k(x,\tilde{x})]\right) \\
&= \Delta - 2\alpha\Delta + \left(\alpha^2\|f^*\|^2 - 2\alpha^2\langle f^*, \mu\rangle + \alpha^2\mathbb{E}[k(x,\tilde{x})]\right) + \alpha^2\Delta \\
&= \Delta - 2\alpha\Delta + \alpha^2\|f^* - \mu\|^2 + \alpha^2\Delta \\
&= \alpha^2\left(\Delta + \|f^* - \mu\|^2\right) - 2\alpha\Delta + \Delta.
\end{aligned}$$

Consequently, we have

$$\Delta_\alpha - \Delta = \alpha^2 \left[\Delta + \|f^* - \mu\|^2\right] - 2\alpha\Delta.$$

This is non-positive where

$$\alpha \in \left[0, \frac{2\Delta}{\Delta + \|f^* - \mu\|^2}\right], \tag{1}$$

and minimized at

$$\alpha_* := \frac{\Delta}{\Delta + \|f^* - \mu\|^2} .$$

This completes the proof. ■

As we can see from (1), there is a range of α for which a non-positive $\Delta_\alpha - \Delta$ is guaranteed. Moreover, the value of α is not necessarily less than 1. To see this, recall that $\widehat{\mu}_\alpha = \alpha f^* + (1-\alpha)\widehat{\mu}$. The distance from $\widehat{\mu}_\alpha$ to the true mean is

$$\Delta_\alpha = \alpha^2 \left[\Delta + \|f^* - \mu\|^2\right] - 2\alpha\Delta + \Delta.$$

When $\alpha = 0$, this distance is Δ. When $\alpha = 2$, the distance is $\|f^* - \mu\|^2 + \Delta$, so if we guess was $f^* = \mu$ then these distances would be exactly the same. We can think of the optimal solution

$$\alpha^* = \frac{\Delta}{\Delta + \|f^* - \mu\|^2}$$

as being the midpoint along a line of "close" solutions which gives the lowest error Δ_α, but there is no reason we cannot move further along this line up until $2\alpha^*$.

3 Kernel Mean Shrinkage Estimator

We give a detailed derivation of both simple kernel mean shrinkage estimator (S-KMSE) and flexible kernel mean shrinkage estimator (F-KMSE). Firstly, note that the loss we define in Section 2 is given by

$$\ell(\mu, g) := \|\mu - g\|^2_{\mathcal{H}} = \|\mathbb{E}[\phi(x)] - g\|^2_{\mathcal{H}} = \mathbb{E}_{xx'}k(x,x') - 2\mathbb{E}_x g(x) + \|g\|^2. \tag{2}$$

By Jensen's inequality, we can upper bound (2) by the loss functional

$$\|\mathbb{E}[\phi(x)] - g\|^2_{\mathcal{H}} \leq \mathbb{E}\|\phi(x) - g\|^2_{\mathcal{H}} =: \mathcal{E}(g). \tag{3}$$

But actually,

$$\mathcal{E}(g) = \mathbb{E}_x k(x,x) - 2\mathbb{E}_x g(x) + \|g\|^2 .$$

Thus, 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. In this paper, we formulate the problem in term of the loss functional (3) as it simplifies the analysis of leave-one-out cross-validation score.

Given an i.i.d. sample $x_1, x_2, \ldots, x_n$, the KMSE can be obtained by minimizing the following loss functional

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

Different choices of $\Omega(\cdot)$ lead to different estimators, as outlined below.

3.1 Simple Shrinkage

By representer theorem, the solution of (4) can be written as $g = \sum_{i=1}^n \beta_i\phi(x_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. Moreover, the S-KMSE uses $\Omega(g) = \|g\|^2_{\mathcal{H}}$. Substituting both $g = \sum_{i=1}^n \beta_i\phi(x_i)$ and $\Omega(\|g\|) = \|g\|^2$ into (4) yields

$$\widehat{\mathcal{E}}_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\left\|\phi(x_i) - \sum_{j=1}^{n}\beta_j\phi(x_j)\right\|^2_{\mathcal{H}} + \frac{\lambda}{2}\left\|\sum_{j=1}^{n}\beta_j\phi(x_j)\right\|^2_{\mathcal{H}}. \tag{5}$$

We can write (5) in term of the kernel function as

$$\begin{aligned}
\widehat{\mathcal{E}}_\lambda(\boldsymbol{\beta}) &= \frac{1}{2n}\sum_{i=1}^{n}\left[k(x_i,x_i) - 2\sum_{j=1}^{n}\beta_j k(x_j,x_i) + \sum_{j=1}^{n}\sum_{k=1}^{n}\beta_j\beta_k k(x_j,x_k)\right] + \frac{\lambda}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} \\
&= \frac{1}{2n}\sum_{i=1}^{n}k(x_i,x_i) - \frac{1}{n}\sum_{i,j=1}^{n}\beta_j k(x_j,x_i) + \frac{1}{2n}\sum_{i,j,k=1}^{n}\beta_j\beta_k k(x_j,x_k) + \frac{\lambda}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} \\
&= \frac{1}{2n}\mathrm{trace}(\mathbf{K}) - \boldsymbol{\beta}^\top\mathbf{K}\mathbf{1}_n + \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} + \frac{\lambda}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} \\
&= \frac{1}{2n}\mathrm{trace}(\mathbf{K}) - \boldsymbol{\beta}^\top\mathbf{K}\mathbf{1}_n + \frac{1}{2}\boldsymbol{\beta}^\top(\mathbf{K} + \lambda\mathbf{K})\boldsymbol{\beta}
\end{aligned}$$

Taking the derivative of $\widehat{\mathcal{E}}_\lambda(\boldsymbol{\beta})$ w.r.t. the vector $\boldsymbol{\beta}$ and setting it to zero yield the optimal weight vector

$$\boldsymbol{\beta} = \left(\frac{1}{1+\lambda}\right)\mathbf{1}_n.$$

Consequently, the shrinkage estimator of the kernel mean is given by

$$\widehat{\mu}_\lambda = \sum_{i=1}^{n}\beta_i\phi(x_i) = \left(\frac{1}{1+\lambda}\right)\widehat{\mu} = (1-\alpha)\widehat{\mu}$$

where $\alpha := \lambda/(1+\lambda) < 1$ and $\widehat{\mu}$ denotes the standard kernel mean estimator.

3.2 Flexible Shrinkage

Using the expansion $g = \sum_{j=1}^{n} \beta_j \phi(x_j)$, the flexible KMSE is obtained by minimizing

$$\widehat{\mathcal{E}}_\lambda(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\left\|\phi(x_i) - \sum_{j=1}^{n}\beta_j\phi(x_j)\right\|_{\mathcal{H}}^2 + \frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta}$$

with respect to the weight vector $\boldsymbol{\beta} \in \mathbb{R}^n$. It can be rewritten in term of the kernel function as

$$\begin{aligned}\widehat{\mathcal{E}}_\lambda(\boldsymbol{\beta}) &= \frac{1}{2n}\text{trace}(\mathbf{K}) - \boldsymbol{\beta}^\top\mathbf{K}\mathbf{1}_n + \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} + \frac{\lambda}{2}\boldsymbol{\beta}^\top\boldsymbol{\beta} \\ &= \frac{1}{2n}\text{trace}(\mathbf{K}) - \boldsymbol{\beta}^\top\mathbf{K}\mathbf{1}_n + \frac{1}{2}\boldsymbol{\beta}^\top(\mathbf{K}+\lambda\mathbf{I})\boldsymbol{\beta}\end{aligned}$$

Taking the derivative of $\widehat{\mathcal{E}}_\lambda(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and setting it to zero yield

$$\begin{aligned}\frac{\partial\widehat{\mathcal{E}}_\lambda}{\partial\boldsymbol{\beta}} = 0 \Rightarrow -\mathbf{K}\mathbf{1}_n + (\mathbf{K}+\lambda\mathbf{I})\boldsymbol{\beta} = 0 \\ (\mathbf{K}+\lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{K}\mathbf{1}_n \\ \boldsymbol{\beta} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n\end{aligned}$$

where $\mathbf{1}_n$ denotes an $n \times 1$ vector whose elements are all $1/n$.

4 Proof of Theorem 2

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 the RKHS $\mathcal{H}$.

Proof. Assume that we know the eigendecomposition $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$ where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ consists of orthogonal eigenvectors of $\mathbf{K}$ such that $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{D} = \text{diag}(\gamma_1, \gamma_2, \ldots, \gamma_n)$ consists of corresponding eigenvalues. Hence, the weights $\boldsymbol{\beta}$ of the F-KMSE is given by

$$\boldsymbol{\beta} = (\mathbf{U}\mathbf{D}\mathbf{U}^\top + \lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n = (\mathbf{U}(\mathbf{D}+\lambda\mathbf{I})\mathbf{U}^\top)^{-1}\mathbf{K}\mathbf{1}_n = \mathbf{U}(\mathbf{D}+\lambda\mathbf{I})^{-1}\mathbf{U}^\top\mathbf{K}\mathbf{1}_n.$$

Consequently,

$$\boldsymbol{\beta} = \sum_{i=1}^{n}\mathbf{u}_i\left(\frac{1}{\gamma_i+\lambda}\right)\mathbf{u}_i^\top\mathbf{K}\mathbf{1}_n. \tag{6}$$

Note also that

$$\mathbf{K}\mathbf{1}_n = \left[\frac{1}{n}\sum_{j=1}^{n}k(x_j, x_1), \ldots, \frac{1}{n}\sum_{j=1}^{n}k(x_j, x_n)\right]^\top = [\langle\widehat{\mu}, \phi(x_1)\rangle, \ldots, \langle\widehat{\mu}, \phi(x_n)\rangle]^\top.$$

Thus, we can rewrite (6) as

$$\begin{aligned}\boldsymbol{\beta} &= \sum_{i=1}^{n} \mathbf{u}_i \left(\frac{1}{\gamma_i+\lambda}\right) \sum_{j=1}^{n} u_{ij} \langle \widehat{\mu}, \phi(x_j)\rangle \\ &= \sum_{i=1}^{n} \mathbf{u}_i \left(\frac{\sqrt{\gamma_i}}{\gamma_i+\lambda}\right) \left\langle \widehat{\mu}, \frac{1}{\sqrt{\gamma_i}} \sum_{j=1}^{n} u_{ij}\phi(x_j)\right\rangle\end{aligned}$$

It follows from the correspondence between the eigenvectors of kernel matrix $\mathbf{K}$ and covariance matrix $\widehat{\mathbf{C}}_{xx}$ that $\mathbf{v}_i = (1/\sqrt{\gamma_i})\sum_j u_{ij}\phi(x_j)$ where $\mathbf{v}_i$ is the ith eigenvector of the covariance matrix. Consequently, we have

$$\left\langle \widehat{\mu}, \frac{1}{\sqrt{\gamma_i}} \sum_{j=1}^{n} u_{ij}\phi(x_j)\right\rangle = \langle \widehat{\mu}, \mathbf{v}_i\rangle \tag{7}$$

In words, (7) is a projection of the standard kernel mean embedding onto the eigenvector $\mathbf{v}_i$. Using this representation, the shrinkage estimate of the F-KMSE given by the weights $\boldsymbol{\beta}$ becomes

$$\widehat{\mu}_\lambda = \sum_{j=1}^{n} \left[\sum_{i=1}^{n} \mathbf{u}_i \left(\frac{\sqrt{\gamma_i}}{\gamma_i+\lambda}\right) \langle \widehat{\mu}, \mathbf{v}_i\rangle\right]_j \phi(x_j).$$

Applying the same trick, we can write the F-KMSE estimate entirely in term of eigenvectors of the covariance matrix $\widehat{\mathbf{C}}_{xx}$ as

$$\begin{aligned}\widehat{\mu}_\lambda &= \sum_{j=1}^{n} \phi(x_j) \sum_{i=1}^{n} u_{ij} \left(\frac{\sqrt{\gamma_i}}{\gamma_i+\lambda}\right) \langle \widehat{\mu}, \mathbf{v}_i\rangle \\ &= \sum_{i=1}^{n} \left(\frac{\sqrt{\gamma_i}}{\gamma_i+\lambda}\right) \langle \widehat{\mu}, \mathbf{v}_i\rangle \sum_{j=1}^{n} u_{ij}\phi(x_j) \\ &= \sum_{i=1}^{n} \left(\frac{\gamma_i}{\gamma_i+\lambda}\right) \langle \widehat{\mu}, \mathbf{v}_i\rangle\, \mathbf{v}_i\end{aligned}$$

Since $\lambda > 0$, we have that $\gamma_i/(\gamma_i+\lambda) < 1$. This completes the proof. ■

5 Proof of Theorem 3

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

Proof. Note that the leave-one-out cross-validation score for the S-KMSE is

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

which can be simplified further as

$$\begin{aligned}
LOOCV(\alpha) &= \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{n}{n-1}(1-\alpha)\widehat{\mu} - \frac{1-\alpha}{n-1}\phi(x_i) - \phi(x_i)\right\|_{\mathcal{H}}^2 \\
&= \left\|\frac{n}{n-1}(1-\alpha)\widehat{\mu}\right\|_{\mathcal{H}}^2 - \frac{2}{n}\left\langle \sum_{i=1}^{n}\frac{n-\alpha}{n-1}\phi(x_i), \frac{n}{n-1}(1-\alpha)\widehat{\mu}\right\rangle \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left\|\frac{n-\alpha}{n-1}\phi(x_i)\right\|_{\mathcal{H}}^2 \\
&= \frac{n^2(1-\alpha)^2}{(n-1)^2}\|\widehat{\mu}\|^2 - \left(\frac{2}{n}\right)\left(\frac{(n-\alpha)n}{n-1}\right)\left(\frac{n(1-\alpha)}{n-1}\right)\|\widehat{\mu}\|^2 \\
&\quad + \frac{1}{n}\left(\frac{n-\alpha}{n-1}\right)^2\sum_{i=1}^{n}k(x_i,x_i) \\
&= \left(\frac{n^2(1-\alpha)^2}{(n-1)^2} - \frac{2n(n-\alpha)(1-\alpha)}{(n-1)^2}\right)\|\widehat{\mu}\|^2 \\
&\quad + \frac{(n-\alpha)^2}{n(n-1)^2}\sum_{i=1}^{n}k(x_i,x_i)
\end{aligned}$$

Let $\rho := \frac{1}{n^2}\sum_{i,j=1}^{n}k(x_i,x_j)$ and $\varrho := \frac{1}{n}\sum_{i=1}^{n}k(x_i,x_i)$. Then, the leave-one-out score becomes

$$LOOCV(\alpha) = \frac{1}{(n-1)^2}\left\{(-n^2+\alpha^2n^2+2\alpha n-2\alpha^2 n)\rho + (n^2-2\alpha n+\alpha^2)\varrho\right\}$$

Taking the derivative of $LOOCV(\alpha)$ with respect to α and setting it to zero yield

$$\alpha_* = \frac{\varrho-\rho}{(n-2)\rho+\varrho/n},$$

Since the parameter α is given by $\alpha = \lambda/(1+\lambda)$, it follows that

$$\lambda_* = \frac{\varrho-\rho}{(n-1)\rho+\varrho/n-\varrho}$$

as required. ■

6 Proof of Theorem 4

In this section we adopt the approach similar to the one presented in P. J. Green (1994) for ridge regression problem. 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

$$\frac{1}{2(n-1)}\sum_{j\neq i}\left\|\phi(x_j)-\sum_{k\neq i}\beta_k\phi(x_k)\right\|_{\mathcal{H}}^2+\frac{\lambda}{2}\|\boldsymbol{\beta}\|^2.$$

We will measure the quality of $\widehat{\mu}_{\lambda}^{(-i)}$ by how well it approximates $\phi(x_i)$. The overall quality of the estimate can be quantified by the cross-validation score function

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

Note that the vector $\boldsymbol{\beta}^{(-i)}$ has length $n-1$, whereas the original vector $\boldsymbol{\beta}$ has length n. To simplify the following analysis, we will assume that $\boldsymbol{\beta}^{(-i)}$ has length n with $\beta_i=0$. Note that this representation does not alter the leave-one-out estimate $\widehat{\mu}_{\lambda}^{(-i)}$.

Theorem 4. The LOOCV score of F-KMSE satisfies

$$LOOCV(\lambda)=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}^{\top}\mathbf{K}-\mathbf{K}_{\cdot i})^{\top}\mathbf{C}_{\lambda}(\boldsymbol{\beta}^{\top}\mathbf{K}-\mathbf{K}_{\cdot i})$$

where $\boldsymbol{\beta}$ is the weight vector calculated from the full dataset with the shrinkage parameter λ and $\mathbf{C}_{\lambda}=(\mathbf{K}-\frac{1}{n}\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{K})^{-1}\mathbf{K}(\mathbf{K}-\frac{1}{n}\mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{K})^{-1}$.

Note that the leave-one-out cross-validation score in Theorem 4 does not depend on the leave-one-out solution $\boldsymbol{\beta}_{\lambda}^{(-i)}$, but depends only on the non-leave-one-out solution $\boldsymbol{\beta}_{\lambda}$. Consequently, the overall score can be computed efficiently.

Proof of Thorem 4. To prove Theorem 4, we first show that the leave-one-out solution $\boldsymbol{\beta}_{\lambda}^{(-i)}$ can be obtained via the standard formulation with modified target vector.

Lemma 2. *For fixed λ and i, let $\boldsymbol{\beta}^{(-i)}$ denote the vector with components $\beta_j^{(-i)}$ for $j\neq i$. Let us define a vector $\Phi^*=[\phi(x_1),\dots,\phi(x_{i-1}),\widehat{\mu}_{\lambda}^{(-i)},\phi(x_{i+1}),\dots,\phi(x_n)]^{\top}$ and a matrix $\mathbf{B}^*_{ml}=\langle\phi(x_m),\Phi^*_l\rangle_{\mathcal{H}}$. Then $\boldsymbol{\beta}^{(-i)}=(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{B}^*\mathbf{1}_n$.*

Proof. For any vector $\boldsymbol{\beta}$,

$$\begin{aligned}\sum_{j=1}^{n}\left\|\Phi^*_j-\sum_{k=1}^{n}\beta_k\phi(x_k)\right\|_{\mathcal{H}}^2+\lambda\|\boldsymbol{\beta}\|^2 &\geq \sum_{j\neq i}\left\|\Phi^*_j-\sum_{k=1}^{n}\beta_k\phi(x_k)\right\|_{\mathcal{H}}^2+\lambda\|\boldsymbol{\beta}\|^2\\ &\geq \sum_{j\neq i}\left\|\Phi^*_j-\sum_{k=1}^{n}\beta_k^{(-i)}\phi(x_k)\right\|_{\mathcal{H}}^2+\lambda\|\boldsymbol{\beta}^{(-i)}\|^2\\ &= \sum_{j=1}^{n}\left\|\Phi^*_j-\sum_{k=1}^{n}\beta_k^{(-i)}\phi(x_k)\right\|_{\mathcal{H}}^2+\lambda\|\boldsymbol{\beta}^{(-i)}\|^2\end{aligned}$$

by the definition of $\boldsymbol{\beta}^{(-i)}$ and the fact that $\Phi^*_i = \widehat{\mu}^{(-i)}_\lambda$. It follows that $\boldsymbol{\beta}^{(-i)}$ is the minimizer of $\sum_j \|\Phi^*_j - \sum_k \beta_k \phi(x_k)\|^2_{\mathcal{H}} + \lambda\|\boldsymbol{\beta}\|^2$ so that $\boldsymbol{\beta}^{(-i)} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{B}^*\mathbf{1}_n$, as required. ■

As we can see, the resulting formulation of $\boldsymbol{\beta}^{(-i)}$ in Lemma 2 depends on the leave-one-out solution $\widehat{\mu}^{(-i)}_\lambda$ which in turn requires a knowledge of $\boldsymbol{\beta}^{(-i)}$. As a result, we cannot use this formulation to compute $\boldsymbol{\beta}^{(-i)}$ in practice. However, it will be very useful as an intermediate step in deriving the leave-one-out cross-validation score.

In the following, we will write $\mathbf{A}$ for $(\mathbf{K}+\lambda\mathbf{I})^{-1}$ throughout. By virtue of Lemma 2, we can write an expression for the deleted residual $\phi(x_i) - \widehat{\mu}^{(-i)}_\lambda$ as

$$\begin{aligned}
\widehat{\mu}^{(-i)}_\lambda - \phi(x_i) &= \sum_{j=1}^{n} \beta^{(-i)}_j \phi(x_j) - \phi(x_i) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{m=1}^{n} \{\mathbf{AB}^*\}_{jm}\, \phi(x_j) - \phi(x_i) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{m\neq i} \{\mathbf{AK}\}_{jm}\, \phi(x_j) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\mathbf{B}^*_{li}\phi(x_j) - \phi(x_i) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{m\neq i} \{\mathbf{AK}\}_{jm}\, \phi(x_j) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \widehat{\mu}^{(-i)}_\lambda\rangle\phi(x_j) - \phi(x_i) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{m=1}^{n} \{\mathbf{AK}\}_{jm}\, \phi(x_j) - \phi(x_i) \\
&\quad - \frac{1}{n}\sum_{j=1}^{n} \{\mathbf{AK}\}_{ji}\, \phi(x_j) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \widehat{\mu}^{(-i)}_\lambda\rangle\phi(x_j) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{m=1}^{n} \{\mathbf{AK}\}_{jm}\, \phi(x_j) - \phi(x_i) \\
&\quad - \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \phi(x_i)\rangle\phi(x_j) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \widehat{\mu}^{(-i)}_\lambda\rangle\phi(x_j) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{m=1}^{n} \{\mathbf{AK}\}_{jm}\, \phi(x_j) - \phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \widehat{\mu}^{(-i)}_\lambda - \phi(x_i)\rangle\phi(x_j) \\
&= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \widehat{\mu}^{(-i)}_\lambda - \phi(x_i)\rangle\phi(x_j)
\end{aligned}$$

Denote the deleted residual $\widehat{\mu}^{(-i)}_\lambda - \phi(x_i)$ by $\Delta^{(-i)}_\lambda$. Then, the above equation can be rewritten as

$$\Delta^{(-i)}_\lambda = \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\langle\phi(x_l), \Delta^{(-i)}_\lambda\rangle\phi(x_j). \qquad (8)$$

Since the deleted residual $\Delta_\lambda^{(-i)}$ lies in the subspace spanned by the samples $\phi(x_1),\dots,\phi(x_n)$, we may write

$$\Delta_\lambda^{(-i)} = \sum_{k=1}^{n} \xi_k \phi(x_k)$$

for some $\boldsymbol{\xi} \in \mathbb{R}^n$. Substituting back into (8) yields

$$\begin{aligned}
\sum_{k=1}^{n} \xi_k \phi(x_k) &= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j,l} \mathbf{A}_{jl} \langle \phi(x_l), \sum_{k=1}^{n} \xi_k \phi(x_k)\rangle \phi(x_j) \\
&= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j,l} \mathbf{A}_{jl} \sum_{k=1}^{n} \xi_k \langle \phi(x_l), \phi(x_k)\rangle \phi(x_j) \\
&= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j,l} \mathbf{A}_{jl} \sum_{k=1}^{n} \xi_k \mathbf{K}_{lk} \phi(x_j) \\
&= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} \mathbf{A}_{jl}\mathbf{K}_{lk}\xi_k \phi(x_j) \\
&= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \{\mathbf{AK}\}_{jk}\, \xi_k \phi(x_j) \\
&= \widehat{\mu}_\lambda - \phi(x_i) + \frac{1}{n}\sum_{j=1}^{n} \{\mathbf{AK}\boldsymbol{\xi}\}_j\, \phi(x_j)
\end{aligned}$$

By taking the inner product on both sides of the equation with respect to the samples $\phi(x_1),\dots,\phi(x_n)$, the optimal $\boldsymbol{\xi}$ can be obtained by solving the system of equations:

$$\begin{aligned}
\mathbf{K}\boldsymbol{\xi} &= \boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i} + \frac{1}{n}\mathbf{KAK}\boldsymbol{\xi} \\
(\mathbf{K} - \frac{1}{n}\mathbf{KAK})\boldsymbol{\xi} &= \boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i} \\
\boldsymbol{\xi} &= (\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}(\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i}),
\end{aligned}$$

where $\mathbf{K}_{\cdot i}$ denotes the ith column of matrix $\mathbf{K}$. Consequently, the leave-one-out cross-validation score for the sample x_i can be computed by

$$\begin{aligned}
\left\|\Delta_\lambda^{(-i)}\right\|_{\mathcal{H}}^2 = \boldsymbol{\xi}^\top \mathbf{K}\boldsymbol{\xi} &= (\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i})^\top (\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}\mathbf{K}(\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}(\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i}) \\
&= (\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i})^\top \mathbf{C}_\lambda (\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i})
\end{aligned}$$

where $\mathbf{C}_\lambda = (\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}\mathbf{K}(\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}$. Hence, we have the score over full dataset

$$LOOCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\|\Delta_\lambda^{(-i)}\right\|_{\mathcal{H}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i})^\top \mathbf{C}_\lambda (\boldsymbol{\beta}^\top \mathbf{K} - \mathbf{K}_{\cdot i})$$

as required. ■

6.1 Efficient Calculation of $\mathbf{C}_\lambda$

The naïve calculation of $\mathbf{C}_\lambda$ can be computationally expensive. Fortunately, it can be simplified by the eigendecomposition of $\mathbf{K}$ as follows:

$$\begin{aligned}
\mathbf{C}_\lambda &= (\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1}\mathbf{K}(\mathbf{K} - \frac{1}{n}\mathbf{KAK})^{-1} \\
&= \mathbf{U}(\mathbf{D} - \frac{1}{n}\mathbf{D}(\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D})^{-1}\mathbf{D}(\mathbf{D} - \frac{1}{n}\mathbf{D}(\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D})^{-1}\mathbf{U}^\top
\end{aligned}$$

Since the last equation only involves the diagonal matrices, it can be computed efficiently. The inversion of the diagonal matrix is just the reciprocal of the diagonal elements. Thus, we can evaluate it in $\mathcal{O}(n)$ operations. Overall, the computational complexity of the leave-one-out cross-validation approximation is only $\mathcal{O}(n^2)$, 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, the hyper-parameters are often selected by a simple grid-based search method. That is, the LOOCV score is evaluated at a set of values of λ on a regular grid with even spacing. Alternatively, as the LOOCV function is a relatively smooth function of the shrinkage parameter λ, the derivative-free method as implemented in `fminsearch` or `fminbnd` routines of the MATLAB optimization toolbox provides simple and efficient way to find an optimal value of λ. Lastly, it is important to note that we cannot use the leave-one-out cross-validation score proposed in the previous section to select kernel parameters because our loss function also depends on the choice of kernel function.

7 Probabilistic View of Kernel Mean Estimation

The kernel mean estimation can be understood probabilistically. First, one should note the difference between *primal form* and *dual form* of the kernel mean estimation. In the primal problem, we consider the following average loss functional:

$$\mathcal{E}_{primal}(g) := \frac{1}{n}\sum_{i=1}^{n} \|\phi(x_i) - g\|^2_{\mathcal{H}} \tag{9}$$

where the desired solution g lies in the RKHS $\mathcal{H}$. Estimating g directly from (9) can be difficult as the RKHS $\mathcal{H}$ is usually high-dimensional, if not infinite, e.g., the RKHS associated with the Gaussian RBF kernel. By representer theorem we can transform the problem (9) into the dual form

$$\mathcal{E}_{dual}(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \left\| \phi(x_i) - \sum_{j=1}^{n} \beta_j \phi(x_j) \right\|^2_{\mathcal{H}}, \tag{10}$$

where we have $g = \sum_{i=1}^n \beta_i \phi(x_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^n$. As a result, the estimation of g is amount to estimating the weight vector $\boldsymbol{\beta}$. We can rewrite the dual form (10) in term of the kernel matrix $\mathbf{K}$ as

$$\mathcal{E}_{dual}(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{K}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{K}\mathbf{1}_n + \frac{1}{n}\mathrm{trace}(\mathbf{K})\,. \tag{11}$$

The standard kernel mean estimator

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

can be obtained as a minimizer of the primal form (9). The corresponding value of $\boldsymbol{\beta}$, i.e., $\boldsymbol{\beta} = \mathbf{1}_n$, is a minimizer of the dual form (10). We assume without loss of generality that the kernel matrix $\mathbf{K}$ is invertible.

One can see in (11) that the dual form is quadratic in $\boldsymbol{\beta}$, which thereby implies that it can be viewed as a negative log-likelihood of some Gaussian distribution over $\boldsymbol{\beta}$. Let $\mathcal{N}(\boldsymbol{\beta};\nu,\Sigma)$ be the Gaussian distribution over $\boldsymbol{\beta}$ with mean ν and covariance matrix Σ. Consequently, we have

$$\begin{aligned}
\mathcal{E}'(\boldsymbol{\beta}) := -\ln\mathcal{N}(\boldsymbol{\beta};\mathbf{1}_n,\mathbf{K}^{-1}) &= -\ln\left[\frac{1}{\sqrt{(2\pi)^n|\mathbf{K}^{-1}|}}\exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{1}_n)^\top\mathbf{K}(\boldsymbol{\beta}-\mathbf{1}_n)\right)\right] \\
&= \ln\sqrt{(2\pi)^n|\mathbf{K}^{-1}|} + \frac{1}{2}(\boldsymbol{\beta}-\mathbf{1}_n)^\top\mathbf{K}(\boldsymbol{\beta}-\mathbf{1}_n) \\
&= \ln\sqrt{(2\pi)^n|\mathbf{K}^{-1}|} + \frac{1}{2}\mathbf{1}_n\mathbf{K}\mathbf{1}_n + \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{K}\mathbf{1}_n \\
&= \frac{1}{2}\boldsymbol{\beta}^\top\mathbf{K}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{K}\mathbf{1}_n + \mathit{const}
\end{aligned}$$

where *const* denotes constant terms that do not depend on $\boldsymbol{\beta}$. It is easy to see that $\mathcal{E}_{dual}(\boldsymbol{\beta})$ and $\mathcal{E}'(\boldsymbol{\beta})$ have the same minimizer. If $\mathbf{K}$ is strictly positive-definite, the minimizer is unique. As a result, the weight $\boldsymbol{\beta}$ of the standard kernel mean estimator can be consdiered as a maximum-likelihood estimate of the probability distribution $\mathcal{N}(\boldsymbol{\beta};\mathbf{1}_n,\mathbf{K}^{-1})$. Note that this distribution differs from the likelihood in the usual sense, i.e., the probability density of the observations given the parameters. Instead, it specifies the probability density of the weight vector $\boldsymbol{\beta}$. In the following we will denote $\mathcal{N}(\boldsymbol{\beta};\mathbf{1}_n,\mathbf{K}^{-1})$ by P_X.

Despite being different from the standard likelihood, the distribution P_X may still be interpreted as a *data-dependent* belief over possible values of $\boldsymbol{\beta}$. Following standard Bayesian formalism, the modelers may want to specify alternative belief over the values of $\boldsymbol{\beta}$. For example,

$$P_M := \mathcal{N}(\boldsymbol{\beta};\mathbf{0},\boldsymbol{\Sigma}).$$

Combining P_X and P_M yields

$$\begin{aligned}
Q := P_X\cdot P_M &= \mathcal{N}(\boldsymbol{\beta};\mathbf{1}_n,\mathbf{K}^{-1})\cdot\mathcal{N}(\boldsymbol{\beta};\mathbf{0},\boldsymbol{\Sigma}) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{1}_n)^\top\mathbf{K}(\boldsymbol{\beta}-\mathbf{1}_n)\right)\exp\left(-\frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}\right)
\end{aligned}$$

$$\propto \quad \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})(\mathbf{K} + \boldsymbol{\Sigma}^{-1})(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right)$$

where $\bar{\boldsymbol{\beta}} = (\mathbf{K} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{K}\mathbf{1}_n$ and this is recognized as the form of Gaussian with mean $\bar{\boldsymbol{\beta}}$ and covariance matrix A^{-1}

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}; \bar{\boldsymbol{\beta}}, A^{-1})$$

where $A = \mathbf{K} + \boldsymbol{\Sigma}^{-1}$. By imposing different prior on $\boldsymbol{\beta}$, we would obtain different mean $\bar{\boldsymbol{\beta}}$. For example, consider when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ where σ^2 specifies the uncertainty of our belief. Then, we have

$$\bar{\boldsymbol{\beta}} = (\mathbf{K} + \sigma^{-2}\mathbf{I})^{-1}\mathbf{K}\mathbf{1}_n$$

which corresponds to the F-KMSE if we set $\lambda = \sigma^{-2}$. Alternatively, one may consider the covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{K}^{-1}$ which reflects covariance structure obtained from the observations. In which case, we have

$$\bar{\boldsymbol{\beta}} = (\mathbf{K} + \sigma^{-2}\mathbf{K})^{-1}\mathbf{K}\mathbf{1}_n = \frac{1}{1+\sigma^{-2}}\mathbf{K}^{-1}\mathbf{K}\mathbf{1}_n = \frac{1}{1+\sigma^{-2}}\mathbf{1}_n$$

which corresponds to the S-KMSE if we set $\lambda = \sigma^{-2}$.

If we think of P_X as a likelihood, then it encodes the dependence of $\boldsymbol{\beta}$ on the observations x_i through the Gram matrix $\mathbf{K}$. For F-KMSE, the prior P_M is independent of the observations. On the other hand, the "prior" of S-KMSE is data-dependent - it is a function of the x_i. Hence, F-KMSE can be written as the product of a prior and a data-dependent likelihood, but S-KMSE cannot. Thus, it is different from standard Bayesian formalism (Rasmussen and Williams 2006; chap. 2.1). The variance term σ^2 specifies the uncertainty of the priors and thus plays similar role as the regularization parameter λ.

8 Shrinkage Centering in Feature Space

In many applications of kernel methods, it is often assumed that the kernel is centered. That is, the feature map of the data in feature space is given by

$$\tilde{\phi}(x) = \phi(x) - \mathbb{E}[\phi(x)].$$

In practice, the feature mean $\mathbb{E}[\phi(X)]$ is approximated using the empirical average $\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$ such that the centered feature map can be written as

$$\tilde{\phi}(x) = \phi(x) - \frac{1}{n}\sum_{i=1}^{n}\phi(x_i).$$

However, it is very difficult to explicitly center the data because the feature space can be high-dimensional, if not infinite. Schölkopf et al. (1998) showed that we can compute the centered kernel in terms of the non-centered kernel alone.

A direct application of our shrinkage estimators is to replace the empirical average in the above formulation by its shrinkage version, i.e.,

$$\tilde{\phi}(x) = \phi(x) - \sum_{i=1}^{n} \beta_i \phi(x_i)$$

and thereby the centered kernel $\mathbf{K}^c$ can be written as

$$\begin{aligned}
\mathbf{K}^c_{ij} &= \left(\phi(x_i) - \sum_{k=1}^{n} \beta_k \phi(x_k)\right)^\top \left(\phi(x_j) - \sum_{k=1}^{n} \beta_k \phi(x_k)\right) \\
&= \phi(x_i)^\top \phi(x_j) - \phi(x_i)^\top \left[\sum_{k=1}^{n} \beta_k \phi(x_k)\right] - \left[\sum_{l=1}^{n} \beta_l \phi(x_l)^\top\right] \phi(x_j) \\
&\quad + \left[\sum_{k=1}^{n} \beta_k \phi(x_k)^\top\right] \left[\sum_{l=1}^{n} \beta_l \phi(x_l)\right] \\
&= \mathbf{K}_{ij} - \boldsymbol{\beta}^\top \mathbf{K}_{\cdot i} - \mathbf{K}_{\cdot j}^\top \boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{\beta}$ is obtained from the shrinkage estimators. Defining an $n \times n$ matrix $\mathbf{B} = [\boldsymbol{\beta}, \boldsymbol{\beta}, \ldots, \boldsymbol{\beta}]$, we can write a compact expression of centering operation as

$$\mathbf{K}^c = \mathbf{K} - \mathbf{B}^\top \mathbf{K} - \mathbf{K}\mathbf{B} + \mathbf{B}^\top \mathbf{K}\mathbf{B}.$$

Consider a set of test points $x_1^*, x_2^*, \ldots, x_m^*$ and define an $m \times n$ test kernel matrix by

$$\mathbf{L}_{ij} = \langle \phi(x_i^*), \phi(x_j) \rangle_{\mathcal{H}}.$$

Thus, the centered test kernel matrix can be similarly obtained as

$$\mathbf{L}^c = \mathbf{L} - \mathbf{B}_t \mathbf{K} - \mathbf{L}\mathbf{B} + \mathbf{B}_t \mathbf{K}\mathbf{B}$$

where $\mathbf{B}_t = [\boldsymbol{\beta}, \boldsymbol{\beta}, \ldots, \boldsymbol{\beta}]^\top$ denotes an $m \times n$ matrix.

9 Covariance-Operator Shrinkage Estimator

We can extend the idea to improving the estimation of cross-covariance operator on the RKHS. It is a foundation to several kernel-based approaches such as kernel PCA, kernel Fisher discriminant analysis, and kernel CCA. The covariance operator can be seen as a mean function in the joint space.

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 positive definite kernel k_X and k_Y (with feature map ϕ and φ). In this section, we will 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. We assume that $\mathbb{E}_X[k_X(X, X)] < \infty$ and $\mathbb{E}_Y[k_Y(Y, Y)] < \infty$.

One can show that 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 $g \in \mathcal{H}_Y$. If X is equal to Y, we obtain 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 $\widehat{\Sigma}_{YX}$ as

$$\widehat{\Sigma}_{YX} := \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \otimes \varphi(y_i) - \widehat{\mu}_X \otimes \widehat{\mu}_Y \tag{12}$$

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 assume that $\tilde{\phi}$ and $\tilde{\varphi}$ are the centered version of the feature map ϕ and φ, respectively. Then, the empirical cross-covariance operator (12) can be rewritten as

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

which can be obtained as a minimizer of the following loss functional:

$$\widehat{\mathcal{E}}(g) := \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \otimes \tilde{\varphi}(y_i) - g \right\|^2_{\mathcal{H}_X \otimes \mathcal{H}_Y}, \quad g \in \mathcal{H}_X \otimes \mathcal{H}_Y\,. \tag{13}$$

Assume that g lies in the subspace spanned by the data, i.e., $g = \sum_{i=1}^{n} \beta_i \tilde{\phi}(x_i) \otimes \tilde{\varphi}(y_i)$. By the inner product property in product space, we have $\langle \tilde{\phi}(x) \otimes \tilde{\varphi}(y), \tilde{\phi}(x') \otimes \tilde{\varphi}(y') \rangle_{\mathcal{H}_X \otimes \mathcal{H}_Y} = \langle \tilde{\phi}(x), \tilde{\phi}(x') \rangle_{\mathcal{H}_X} \langle \widetilde{\varphi(y)}, \widetilde{\varphi(y')} \rangle_{\mathcal{H}_Y} = k_X(x, x') k_Y(y, y')$.

Note that (13) is of the same form as the kernel mean estimator. As a result, we can apply the same analysis throughout.

References

J. Berger. Minimax estimation of location vectors for a wide class of densities. *Annals of Statistics*, 3(6):1318–1328, 1975.

M. E. Bock. Minimax estimators of the mean of a multivariate normal distribution. *Annals of Statistics*, 3(1):209–218, 1975.

B. Efron and C. Morris. Limiting the risk of bayes and empirical bayes estimators–part ii: The empirical bayes case. *J. Am. Stat. Assoc.*, 67(337):130–139, 1972.

B. Efron and C. Morris. Stein's estimation rule and its Competitors–An empirical bayes approach. *J. Am. Stat. Assoc.*, 68(341):117–130, 1973a.

B. Efron and C. Morris. Combining possibly related estimation problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(3):379–421, 1973b.

B. Efron and C. Morris. Data analysis using stein's estimator and its generalizations. *J. Am. Stat. Assoc.*, 70(350):311–319, 1975.

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

H. M. Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Annals of Statistics*, 6(3):473–484, 1978.

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.

B. W. S. P. J. Green. *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman and Hall, 1994.

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.

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.