Journal of Multivariate Analysis 169 (2019) 166–178

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

A class of optimal estimators for the covariance operator in reproducing kernel Hilbert spaces

Yang Zhou [a], Di-Rong Chen [a,b,*], Wei Huang [c]

[a] School of Mathematics and System Science, Beihang University, Beijing 100191, PR China
[b] Department of Mathematics, Wuhan Textile University, Wuhan 430200, PR China
[c] School of Mathematics, Hefei University of Technology, Hefei 230000, PR China

ARTICLE INFO

Article history:
Received 29 January 2018
Available online 17 September 2018

AMS 2010 subject classifications:
62J10
47B32
62C20
62G20

Keywords:
Covariance operator
Minimax lower bound
Rate of convergence
Reproducing kernel Hilbert space
Shrinkage estimator

ABSTRACT

The covariance operator plays an important role in modern statistical methods and is critical for inference. It is most often estimated by the empirical covariance operator. In spite of its simple and appealing properties, however, this estimator can be improved by a class of shrinkage operators. In this paper, we study shrinkage estimation of the covariance operator in reproducing kernel Hilbert spaces. A data-driven shrinkage estimator enjoying desirable theoretical and computational properties is proposed. The procedure is easily implemented and its numerical performance is investigated through simulations. In finite samples, the estimator outperforms the empirical covariance operator, especially when the data dimension is much larger than the sample size. We also show that the rate of convergence in Hilbert–Schmidt norm is of the order $n^{-1/2}$. Furthermore, we establish the minimax optimal rate of convergence over suitable classes of probability measures and demonstrate that these shrinkage operators are all minimax rate-optimal.

1. Introduction

As a nonlinear tool for data analysis, the kernel method has been widely used in many applications. In nonlinear component analysis, including kernel principal component analysis [2,22], kernel canonical correlation analysis [1,3,6] and kernel Fisher discriminant analysis [14,32], data are represented as functions or elements in a reproducing kernel Hilbert space (RKHS). In predictive learning tasks, kernel mean embedding is used for solving classification [18,23] and regression [27] problems. In statistical hypothesis testing of homogeneity [10], independence [12] and conditional independence [9], as well as in kernel-based dimensionality reduction for supervised learning [7] and regression [8], kernels offer a linear approach to deal with higher order statistics [25]. All of these methods depend heavily on the covariance operator in RKHS and thus its estimation is one of the most basic issues in practice.

It is assumed that $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ is an RKHS with a continuous reproducing kernel $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ defined over a separable topological space $\mathcal{X}$. A kernel mean of a probability distribution P over $\mathcal{X}$ is defined by a Bochner integral, viz.

$$\mu_P = \int_{\mathcal{X}} K(\cdot, x)dP(x) \in \mathcal{H}_K; \tag{1}$$

see Chapter 1 in [4]. A sufficient condition for the existence of μ_P is $\int_{\mathcal{X}} \sqrt{K(x,x)}dP < \infty$; see [24]. The embedding μ_P represents the expectation of functions in $\mathcal{H}_K$ in the form of an inner product $\mathrm{E}f(X) = \langle\mu_P, f\rangle_{\mathcal{H}}$ by the reproducing property.

* Corresponding author at: School of Mathematics and System Science, Beihang University, Beijing 100191, PR China.
E-mail addresses: yangz91@buaa.edu.cn (Y. Zhou), drchen@buaa.edu.cn (D.-R. Chen), whuang@hfut.edu.cn (W. Huang).

https://doi.org/10.1016/j.jmva.2018.09.003
0047-259X/

Moreover, under the stronger condition $\int_{\mathcal{X}} K(x,x)dP < \infty$, the covariance operator associated with P on $\mathcal{H}_K$, given, for all $f \in \mathcal{H}_K$, by

$$\Sigma_P f = \mathrm{E}[\{f(X) - \mathrm{E}f(X)\}\{K(\cdot, X) - \mu_P\}],$$

is well defined and is a Hilbert–Schmidt (HS) operator; see [11]. Given a random sample $X_1, \ldots, X_n$ from P, the empirical covariance operator has often been used as a standard estimator of Σ_P. It is given by

$$\Sigma_{P_n}(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\langle K_{X_i} - \mu_{P_n}, \cdot\rangle_{\mathcal{H}}(K_{X_i} - \mu_{P_n}), \tag{2}$$

where $\mu_{P_n} = (K_{X_1} + \cdots + K_{X_n})/n$ is the empirical average of μ_P and $K_x = K(\cdot, x)$ for any fixed $x \in \mathcal{X}$.

It is well known that the empirical covariance operator has low bias but that its high variance leads to a high mean square error (MSE). For this reason, many attempts have been made to improve on it using shrinkage methods. The idea of shrinkage is to make a trade-off between bias and variance by modifying the standard empirical estimator. A basic strategy is to increase bias and decrease the variance simultaneously in such a way as to reduce the MSE. In [19], two shrinkage estimators called simple covariance shrinkage estimator (SCOSE) and flexible covariance shrinkage estimator (FCOSE) are proposed based on regularization techniques. Furthermore, Ramdas et al. [21] employed them to the independence testing problem and improved on the power of the associated independence test. These shrinkage estimators perform favorably in finite samples, especially when the data dimension is much larger than the sample size. In high-dimensional statistics, recent research in shrinkage methods can be found in [5,16,28].

In this paper, we show from theoretical and empirical considerations that the standard estimator is suboptimal in the sense of MSE and that there exist estimators that can improve upon it. These estimators are constructed based on shrinkage methods. More specifically, a linear combination of a preassigned HS operator and the standard estimator is considered. The preassigned operator reflects our prior knowledge of covariance and determines the direction of shrinkage. Under the HS-norm criterion, the optimal shrinkage intensity that determines the best linear combination depends on the unknown covariance operator and it should be estimated from data. With a simple but effective estimation of the optimal shrinkage intensity, a data-driven shrinkage estimator is proposed. This is one of this paper's main contributions. The simulations in Section 4 illustrate that the proposed estimator has satisfactory performance and is suitable in small-sample settings.

Another contribution of this paper is to establish the asymptotic convergence rate of the shrinkage estimator in HS-norm. Various shrinkage estimators have been introduced which behave well in practice [16,17,28]. However, not much is known about their theoretical properties. To comprehend the effectiveness of shrinkage methods, we establish a uniform upper bound at the rate of $n^{-1/2}$ for the shrinkage estimator under a mild assumption on P. This implies that the proposed estimator converges to Σ_P at the same rate as Σ_{P_n}. Moreover, we study the minimax rate of convergence for estimating the covariance operator in a reproducing kernel Hilbert space. Under mild conditions on K, we show that $n^{-1/2}$ is the optimal minimax rate by deriving minimax lower bounds over several classes of Borel probability measures. An interesting aspect of this result is that the minimax rate is independent of the smoothness of the kernel K and the density of P. These results show that the standard estimator and the proposed estimator are both rate-optimal.

The rest of the paper is organized as follows. In Section 2, we construct a class of shrinkage operators for Σ_P and propose a completely data-driven shrinkage estimator. In Section 3, we establish a uniform upper bound for this estimator and derive lower bounds over several classes of distributions, implying that the established bound is in fact rate sharp. In Section 4, the performance of the proposed estimator is illustrated in some simulated examples.

In what follows, the subscript P in Σ_P and μ_P will be dropped if the distribution is not taken into consideration. For simplicity, the HS-norm or $\mathcal{H}_K$-norm is denoted $\|\cdot\|$ when there is no ambiguity, and K_{X_i} is simplified as K_i.

2. Shrinkage estimator

To construct an estimator of the covariance operator, it is helpful to treat it as the expectation of a rank-1 operator $\Sigma = \mathrm{E}\{(K_X - \mu)\otimes(K_X - \mu)\}$ where, for $u, v \in \mathcal{H}_K$, $u \otimes v$ is the rank-1 operator defined by $u \otimes v(f) = \langle u, f\rangle v$ for all $f \in \mathcal{H}_K$. It is easily seen that $u \otimes v$ is an HS operator with $\|u \otimes v\|_{HS} = \|u\|_{\mathcal{H}} \times \|v\|_{\mathcal{H}}$; see [11]. With this notation, the empirical covariance operator in (2) can be rewritten as

$$\Sigma_{P_n} = \frac{1}{n}\sum_{i=1}^{n}(K_i - \mu_{P_n})\otimes(K_i - \mu_{P_n}).$$

As Σ_{P_n} is a biased estimator, it is sometimes replaced by the sample covariance operator $\widetilde{\Sigma} = n\Sigma_{P_n}/(n-1)$.

Generally speaking, a shrinkage estimator is a combination of an estimator with low bias but high variance and another estimator with high bias but low variance. As $\widetilde{\Sigma}$ is an unbiased estimator, given a deterministic HS operator Σ^* on $\mathcal{H}_K$ and a shrinkage parameter α, we consider the shrinkage estimator defined, for all $\alpha \in \mathbb{R}$, by

$$\widehat{\Sigma}_\alpha = \alpha\Sigma^* + (1-\alpha)\widetilde{\Sigma}.$$

The choice of Σ^* is arbitrary but independent of the sample. It is obvious that this shrinkage estimator pulls the raw estimator $\widetilde{\Sigma}$ toward Σ^* by an amount specified by α. If $\alpha = 0$, then $\widehat{\Sigma}_\alpha = \widetilde{\Sigma}$; if $\alpha = 1/n$ and $\Sigma^* = \mathbf{0}$, $\widehat{\Sigma}_\alpha$ reduces to Σ_{P_n}.