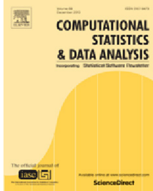
Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



Canonical kernel dimension reduction

Chenyang Tao [a,b], Jianfeng Feng [a,b,c,*]

[a] Centre for Computational Systems Biology and School of Mathematical Sciences, Fudan University, Shanghai, 200433, PR China
[b] Department of Computer Science, Warwick University, Coventry, UK
[c] School of Life Science and the Collaborative Innovation Center for Brain Science, Fudan University, Shanghai, 200433, PR China

HIGHLIGHTS

- A new sufficient dimension reduction method based on kernel canonical functions.
- This new method is distribution free and highly scalable.
- We give theoretical proof of the sufficient dimension reduction property.
- We present efficient algorithms and discuss the choice of loss function.
- Extensive experiments demonstrate its advantage over existing approaches.

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form 10 June 2016
Accepted 4 October 2016
Available online xxxx

Keywords:
Canonical correlation analysis
Canonical functions
Kernel dimension reduction
Krylov subspace
Sufficient dimension reduction
Reproducing kernel Hilbert space

ABSTRACT

A new kernel dimension reduction (KDR) method based on the gradient space of canonical functions is proposed for sufficient dimension reduction (SDR). Similar to existing KDR methods, this new method achieves SDR for arbitrary distributions, but with more flexibility and improved computational efficiency. The choice of loss function in cross-validation is discussed, and a two-stage screening procedure is proposed. Empirical evidence shows that the new method yields favorable performance, both in terms of accuracy and scalability, especially for large and more challenging datasets compared with other distribution-free SDR methods.

1. Introduction

In the era of big data, supervised dimension reduction serves as an invaluable tool to make the best use of the high-dimensional datasets by casting them onto some lower dimensional manifolds with minimum loss of relevant information. The task is to seek a low-dimensional embedding $Z \in \mathbb{R}^d$ of some high-dimensional vector $X \in \mathbb{R}^p$ using information from some auxiliary variable Y, which in most cases is a $\mathbb{R}^q$ vector but can also be more abstract objects such as graphs, texts, *etc*. Popular methods to achieve this task include canonical correlation analysis, partial least square, and LASSO, among others.

One particular research direction is the so-called sufficient dimension reduction (SDR), where a low-dimension representation Z of X that fully captures the conditional distribution of Y given X, i.e., $\mathbb{P}(Y|Z) = \mathbb{P}(Y|X)$, is identified. For

* Corresponding author at: Centre for Computational Systems Biology and School of Mathematical Sciences, Fudan University, Shanghai, 200433, PR China.
E-mail address: jianfeng64@gmail.com (J. Feng).

http://dx.doi.org/10.1016/j.csda.2016.10.003

computational reasons, Z is usually restricted to linear combinations of X, while not prohibiting other forms (Wang et al., 2014). Since the seminal paper of sliced inverse regression (SIR) (Li, 1991), SDR has been extensively studied (Cook and Ni, 2005; Li and Dong, 2009; Ma and Zhu, 2013). In current studies, SDR is approached in three ways: inverse regression, forward regression and joint approach. Inverse regression focuses on the distribution of X given Y, and popular methods in this category include SIR (Li, 1991), sliced average variance estimator (Cook and Weisberg, 1991) and principal Hessian direction (Li, 1992). While these methods are computationally cheap, they depend on such strong assumptions as elliptical distribution of X. Average derivative estimation (Härdle and Stoker, 1989; Samarov, 1993), minimum average variance estimation (Xia et al., 2002) and sliced regression (Wang and Xia, 2008) are examples of forward regression, which focuses on the distribution of Y, given X. They are free of restrictive probability assumptions, yet suffer from heavy computational burden as a result of the nonparametric estimation procedures involved. The joint approach, including methods such as those based on Kullback–Leibler divergence (Yin and Cook, 2005; Yin et al., 2008), mutual information (MI) (Suzuki and Sugiyama, 2013; Tangkaratt et al., 2015), Fourier analysis (Zhu and Zeng, 2006), integral transforms (Zeng and Zhu, 2010), or canonical dependency (Fung et al., 2002; Karasuyama and Sugiyama, 2012), all focus on exploiting the joint distribution of (X, Y).

The pioneering works of Fukumizu have produced kernel dimension reduction (KDR) techniques, such as trace-based kernel dimension reduction (tKDR) (Fukumizu et al., 2004, 2009) and gradient-based kernel dimension reduction (gKDR) (Fukumizu and Leng, 2012). Among other joint approaches, these techniques present solutions to the problem of SDR by embedding probability distributions in the reproducing kernel Hilbert space (RKHS) and exploiting the cross-covariance operators between RKHSs. These methods are also characterized as distribution-free. Apart from its theoretical grounding, KDR also showed very competitive empirical performance. Still, its applications are limited by the heavy computational burden involved, especially for tKDR. Although gKDR is much more efficient than tKDR, it suffers from degenerated accuracy on many benchmark problems when compared to tKDR.

In this work, we describe a novel kernel dimension reduction method that improves upon the accuracy of tKDR, while, at the same time, consuming less computational resources than that of gKDR. Our approach is based on kernel canonical-correlation analysis, and, as such, it is termed as ccaKDR. We prove that the central space is equivalent to the space spanned by the derivative of the canonical functions with nonvanishing eigenvalues in RKHS under mild conditions, and a more scalable linear scaling approximation algorithm is presented. We also present a two-stage screening procedure and discuss the choice of loss function, both topics of pragmatic importance. Empirical evidence reveals that better accuracy and scalability can be expected from ccaKDR compared with other distribution-free alternatives.

The paper is organized as follows. In Section 2, we briefly review the technical tools required, propose ccaKDR and present its theoretical justifications, followed by a discussion of relevant issues. In Section 3, we conduct numerical experiments on both synthetic and real-world data to substantiate the paper. Concluding remarks are given in Section 4. MATLAB code for the algorithms and sample data can be found on the authors' website.

2. CCA-based kernel dimension reduction

2.1. Background

In this section, we briefly review the mathematical tools needed to derive and compute the proposed ccaKDR. We use capital letters $X, Y, \ldots$ to denote random variables, bold font capital letters $\boldsymbol{A}, \boldsymbol{B}, \ldots$ to denote matrices, and use notation $[n]$ for the set $\{1, \ldots, n\}$.

Reproducing kernel Hilbert space (RKHS) has been established as a versatile tool in machine learning, especially for nonlinear problems, with the most prominent examples including support vector machines in classification and regression. We briefly review the basic concepts here. If we denote Ω of some set, then we call a real-valued symmetric function $\kappa(\cdot, \cdot)$ defined on $\Omega \times \Omega$ a positive definite kernel if it satisfies $\sum_{i,j=1}^{n} c_i c_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0$ for any $\{c_i\}_{i=1}^{n} \in \mathbb{R}$ and $\{\boldsymbol{x}_i\}_{i=1}^{n} \in \Omega$ with any $n \geq 0$, and we will hereinafter simply refer to it as a kernel. For such a kernel on Ω, Aronszajn (1950) established that there is a unique Hilbert space $\mathcal{H}$, with its inner product $\langle \cdot, \cdot \rangle$ induced by κ, consisting of functions on Ω such that (i) $\kappa(\cdot, \boldsymbol{x}) \in \mathcal{H}$, (ii) the linear hull of $\{\kappa(\cdot, \boldsymbol{x}) | \boldsymbol{x} \in \Omega\}$ is dense in $\mathcal{H}$, and (iii) for any $\boldsymbol{x} \in \Omega$ and $f \in \mathcal{H}$, $\langle f, \kappa(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}} = f(\boldsymbol{x})$. We note that (iii) is the famous reproducing property and, thus the name reproducing kernel Hilbert space. The representer theorem (Kimeldorf and Wahba, 1970) serves as the foundation of almost all kernel methods, and it basically states that the minimizer of functions in $\mathcal{H}$ of some empirical risk function plus regularization admits the form of a linear combination of $\kappa(\cdot, \boldsymbol{x}_i)$ based on empirical samples $\{\boldsymbol{x}_i\}_i^n$. This equates the optimization on an infinite dimensional search space $\mathcal{H}$ to a finite dimensional search space $\mathbb{R}^n$.

Kernel embedding and cross-covariance operators are theoretical tools developed in recent years for kernel techniques involved with distributions for many statistical problems. Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mu_{\mathcal{X}})$ be the probability measure space for random variable X defined on $\mathcal{X}$ and $(\kappa_{\mathcal{X}}, \mathcal{H}_{\mathcal{X}})$ the measurable kernel and associated RKHS, respectively. A *kernel embedding* of $\mu_{\mathcal{X}}$ with respect to $\kappa_{\mathcal{X}}$ is defined as $\mathbb{E}_{\mu_{\mathcal{X}}}[\kappa(\cdot, X)] \in \mathcal{H}_{\mathcal{X}}$, and if such embedding map from the space of all probability distributions defined on $\mathcal{X}$ to $\mathcal{H}_{\mathcal{X}}$ is injective, then we call the kernel *characteristic*. That is to say for characteristic kernels $\mathbb{E}_{\mu}[\kappa(\cdot, X)] = \mathbb{E}_{\nu}[\kappa(\cdot, X)]$ implies $\mu = \nu$. This is a generalization of the characteristic functions on probability measures, as defined on Euclidean spaces, and popular examples of characteristic kernels include Gaussian kernel and Laplace kernel.