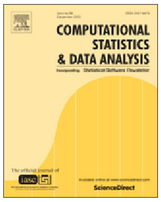
Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



Identifying outliers using multiple kernel canonical correlation analysis with application to imaging genetics

Md. Ashad Alam [a,c,*], Vince D. Calhoun [b], Yu-Ping Wang [a]

[a] *Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118, USA*
[b] *Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM 87131, USA*
[c] *Department of Statistics, Hajee Mohammad Danesh Science and Technology University, Dinajpur 5200, Bangladesh*

ARTICLE INFO

Article history:
Received 23 March 2017
Received in revised form 15 November 2017
Accepted 20 March 2018
Available online xxxx

Keywords:
Multiple kernel CCA
Influence function
Outlier detection
Multimodal datasets
Imaging genetics

ABSTRACT

Identifying significant outliers or atypical objects from multimodal datasets is an essential and challenging issue for biomedical research. This problem is addressed, using the influence function of multiple kernel canonical correlation analysis. First, the influence function (IF) of the kernel mean element, the kernel covariance operator, the kernel cross-covariance operator and kernel canonical correlation analysis (kernel CCA) are studied. Second, an IF of multiple kernel CCA is proposed, which can be applied to multimodal datasets. Third, a visualization method is proposed to detect influential observations of multiple sources of data based on the IF of kernel CCA and multiple kernel CCA. Finally, to validate the method, experiments on both synthesized and imaging genetics data (e.g., SNP, fMRI, and DNA methylation) are performed. To examine the outliers, both the stem-and-leaf display and distribution based technique are used. The performance of the proposed approach is illustrated on 116 candidate regions of interest (ROIs) from the fMRI data of schizophrenia study to identify significant ROIs. The proposed method and two state-of-the-art statistical methods have identified 8, 34, and 10 ROIs, respectively. Based on an online database, the brain mappings of the selected common 7 ROIs indicate the irregular brain regions susceptible to schizophrenia. The results demonstrate that the proposed method is capable of analyzing outliers and the influence of observations, and can be applicable to many other biomedical data which are often high-dimensional and multi-modal.

1. Introduction

Imaging genetics research has essentially focused on discovering unique and co-association effects, but typically ignoring atypical objects in genetics as well as non-genetics variables even when such objects are present. Outliers may be right, but we need to examine for transcription errors, which are commonly made by human operators or a machine. Outliers can also cause difficulties for classical statistical methods (Gogoi et al., 2011). When applying a statistical approach to imaging genetics data containing outliers, results can be deceptive. To overcome this problem, many robust methods have been developed which are less sensitive to outliers. The goal of robust statistics is to use the bulk of the data to identify points deviating from the majority of the data (Huber and Ronchetti, 2009; Hampel et al., 2011; Naser and Hamzah, 2012; Alam et al., 2016). It is well-known that most robust methods are computationally intensive and experience the curse of dimensionality problem. The outliers need to be removed or downweighted prior to fitting non-robust statistical or machine learning approaches (Filzmoser et al., 2008; Oh and Gao, 2009; Roth, 2006).

[*] Corresponding author at: Department of Statistics, Hajee Mohammad Danesh Science and Technology University, Dinajpur 5200, Bangladesh.
E-mail address: malam@tulane.edu (M.A. Alam).

https://doi.org/10.1016/j.csda.2018.03.013
0167-9473/© 2018 Elsevier B.V. All rights reserved.

The incorporation of various unsupervised learning methods into genomic analysis is a rather recent topic. Using the dual representations in problems of supervised and unsupervised learning, the task of learning from multiple data sources is related to kernel-based data integration, which has been actively studied in the last decade (Charpiat et al., 2015; Hofmann et al., 2008; Alam, 2014). Kernel fusion in unsupervised learning has a close connection with unsupervised kernel methods. As unsupervised kernel methods, kernel principal component analysis (kernel PCA) (Schölkopf et al., 1998; Alam and Fukumizu, 2014), kernel canonical correlation analysis (kernel CCA) (Akaho, 2001; Alam and Fukumizu, 2015, 2013), weighted multiple kernel CCA (Yu et al., 2011) have been extensively studied for decades. However, these methods are not robust and thus are sensitive to contaminated data. To apply all of these non-robust methods to genomics data, outliers identification or robust approaches are essential.

Nowadays, influence function (IF) based methods have been used to identify an influence observation. From a statistical perspective, IF reflects the rate of change in a functional upon a small amount of contamination by another distribution (Hampel et al., 1986). Debruyne et al. (2010) have proposed a visualization method for detecting influential observations using the IF of kernel PCA. Filzmoser et al. (2008) developed a method for outlier identification in high dimensions. However, these methods are limited to a single dataset.

Due to the properties of eigen-decomposition, kernel CCA is still a well used method for multiple sources data analysis and integration. Alam et al. (2010) have performed an empirical comparison and sensitivity analysis of robust linear CCA and kernel CCA, giving similar interpretation as kernel PCA (Alam et al., 2010, 2008, 2016; Huang et al., 2009). In addition, Romanazzi (1992) and Alam et al. (2016) have proposed the IF of CCA and kernel CCA but the IF of multiple kernel CCA has not been studied. All of these considerations motivate us to conduct studies on the IF of multiple kernel CCA to identify outliers in imaging genetics datasets.

The contribution of this paper is four-fold. First, we address the IF of kernel mean element (kernel ME), kernel covariance operator (kernel CO), kernel cross-covariance operator (kernel CCO) and kernel CCA. Second, we theoretically derive the IF of multiple kernel CCA, which can be applied for more than two data sets. Third, we propose a visualization method to detect influential observations of multiple datasets. We use the step-and-leaf display and distribution based methods to confirm the outliers or influential observations. Finally, the proposed method is applied to identify outliers in both synthesized and real imaging genetics data (e.g., SNP, fMRI, and DNA methylation), resulting in the detection of significant ROIs in the brain.

The remainder of the paper is organized as follows. In the next section, we provide a brief review of kernel ME, kernel CO, kernel CCO, and its IFs. In Section 3, we discuss kernel CCA and multiple kernel CCA. After a brief review of kernel CCA along with its IF in Section 3.1, we derive the IF of multiple kernel CCA in Section 3.3. The utility of the proposed method is demonstrated by both simulated and real data analysis from an imaging genetics study in Section 4. In Section 5, we summarize our findings and give a perspective for future research. Details derivation can be found in the appendix.

2. Preliminary

Recently, nonparametric statistical inference approaches in reproducing kernel Hilbert space (RKHS) have been widely used. In these approaches, the distribution of a random variable is represented by the kernel ME, which is the mean element of the random feature vector defined by the kernel function, where the relation among variables is expressed by covariance and cross-covariance operators (Gretton et al., 2008; Fukumizu et al., 2008; Song et al., 2008; Kim and Scott, 2012; Gretton et al., 2012).

Let $\mathcal{X}, \mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$ be the unique and joint sample spaces, respectively. Also let F_X, F_Y and F_{XY} be the probability measure on $\mathcal{X}, \mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$, respectively. A symmetric bivariate function, $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined on a space is called a **positive definite kernel** if the Gram matrix $(k(X_i, X_j))_{ij}$ is positive semi-definite for all $i, j \in \{1, 2, \ldots, n\}$, where $X_1, X_2, \ldots, X_n$ are the independent and identically distributed samples from the distribution F_X. Aronszajn (1950) has shown that a positive definite kernel is associated with a Hilbert space, called **reproducing kernel Hilbert space,** $\mathcal{H}_X$. The **feature map** is a mapping $\Phi : X \to \mathcal{H}_X$ and defined as $\Phi(\cdot) = k(\cdot, X), \forall X \in \mathcal{X}$. The vector $\Phi(X) \in \mathcal{H}_X$ is called a **feature vector**. The inner product of two feature vectors can be defined as $\langle \Phi(X), \Phi(X') \rangle_{\mathcal{H}_X} = k(X, X')$ for all $X, X' \in \mathcal{X}$. This is called the **kernel trick**. By the reproducing property, $f(X) = \langle f(\cdot), k(\cdot, X) \rangle_{\mathcal{H}_X}$, with $f \in \mathcal{H}_X$ and the kernel trick, the kernel can evaluate the inner product of any two feature vectors efficiently without knowing an explicit form of either the feature map or the feature vector. In addition, the computational cost does not depend on the dimension of the original space after computing the Gram matrices (Fukumizu and Leng, 2014; Alam and Fukumizu, 2014). In the following sections, we address the basic notations of kernel ME, kernel CO and kernel CCO with their IFs.

2.1. Kernel mean element

Let k_X be a measurable positive definite kernel on $\mathcal{X}$ with $\mathbb{E}_X[\sqrt{k(X, X)}] < \infty$. The **kernel mean element**, $\mathcal{M}_X$, of X on $\mathcal{H}_X$ is an element of $\mathcal{H}_X$ and is defined by the mean of the $\mathcal{H}_X$-valued random variable $k_X(\cdot, X)$,

$$\mathcal{M}_X(\cdot) = \mathbb{E}_X[k_X(\cdot, X)].$$

The kernel mean always exists with arbitrary probability under the assumption that the positive definite kernels are bounded and measurable. By the reproducing property, the kernel ME satisfies the following equality

$$\langle \mathcal{M}_X, f \rangle_{\mathcal{H}_X} = \langle \mathbb{E}_X[k_X(\cdot, X)], f \rangle_{\mathcal{H}_X} = \mathbb{E}_X \langle k_X(\cdot, X), f \rangle_{\mathcal{H}_X} = \mathbb{E}_X[f(X)],$$

for all $f \in \mathcal{H}_X$.

Please cite this article in press as: Alam M.A., et al., Identifying outliers using multiple kernel canonical correlation analysis with application to imaging genetics. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.03.013.