Journal of Multivariate Analysis 134 (2015) 1–18

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Canonical correlation analysis for irregularly and sparsely observed functional data

Hyejin Shin [a], Seokho Lee [b,*]

[a] Bell Labs, Seoul, South Korea
[b] Hankuk University of Foreign Studies, Yongin, South Korea

ARTICLE INFO

Article history:
Received 3 March 2014
Available online 23 October 2014

AMS subject classifications:
62J05
62G20
62G35

Keywords:
Canonical correlation analysis
Functional data analysis
Linear mixed-effects model
Longitudinal data
Reproducing kernel Hilbert spaces

ABSTRACT

Several approaches for functional canonical correlation analysis have been developed to measure the association between paired functional data. However, the existing methods in the literature have been developed for dense and balanced functional data, and they cannot be directly applicable to the situations where the observed curves are recorded in the irregular and sparse fashion. In this paper, we model the associations between paired functional data into a linear mixed-effects model framework by relating two sets of curves using canonical correlation analysis. The proposed approach automatically deals with irregularly or sparsely observed functional data, and brings a new insight into the interpretation of canonical correlation analysis. Numerical studies are carried out to demonstrate finite sample behavior. Two real data applications are provided to illustrate the methodology.

1. Introduction

Functional canonical correlation analysis (FCCA) is a useful tool to quantify the relationship between paired functional data. Several approaches for FCCA have been developed so far. Recent work on FCCA includes [15,8,9,4,3,7,11]. However, the existing methods have been developed for dense and balanced functional data, and they cannot be directly applied or easily extended to the situations where the observed curves are measured at an irregular and sparse set of time points. This is because the cross-correlation structure of two sets of irregularly and sparsely observed functional data is not readily obtainable.

Some approaches, other than FCCA, have been suggested to study the association of paired longitudinal data in the functional-data context. Dubin and Müller [6] introduced the concept of dynamical correlation to measure the dependency between multivariate longitudinal data by treating longitudinal pathways as realizations of a smooth stochastic process. They initially compute the individual correlation between paired smoothed curves at the subject level and then obtain dynamical correlation by averaging the correlations over the subjects in the sample. This approach, however, is not applicable to sparsely observed functional data since it requires the pre-smoothing step to individual curves having a few measurements. Yao et al. [18] studied functional linear regression through principal component analysis to model the relationship between two sets of sparsely observed functional data, and Zhou et al. [19] proposed a mixed-effects model for relationship between two sparsely observed functional data by modeling the associations through principal component scores. The common strategy of these two approaches is that they model the association of two sets of unbalanced

* Corresponding author.
E-mail address: lees@hufs.ac.kr (S. Lee).

http://dx.doi.org/10.1016/j.jmva.2014.10.001

longitudinal data through the principal component scores associated with a few major principal components. PC-based association approach, however, may not be effective in finding association between paired curves, because those principal components are usually sought to have maximal variability of each set of curves, not to represent maximal cross-correlation between two sets of curves. Therefore, there is a possibility that some principal components explaining non-ignorable amount of associations are dropped from the final models, failing to detect the significant modes of covariability. Although a better strategy for functional association is to use canonical correlation analysis (CCA), no functional adaptation of CCA method for sparse functional data is available.

While CCA and its value in statistical analysis have been widely accepted among the audience of statistics and other related fields, CCA does not seem popular in the data analysis as much as principal component analysis (PCA) and factor analysis (FA) do. This is mainly because we are less frequently given the data having two sets of variables rather than the data with which a single-set variable analysis is appropriate. Another reason, that is of our concern, is that CCA itself has the limited usage in the data analysis. CCA is useful in determining the strength of the relationships between two sets of variables, and deriving a set of variable weights (canonical loadings) in the linear combination of one set of variables, which is maximally correlated with the linear combination of the other set of variables. Unlike PCA and FA, CCA machinery does not directly tell us which directions among two sets of variables are co-varying with the maximal correlation in the data space. PCA provides the directional vector (PC loading) that explains the maximal variability. The change in scale in principal component means that the process shows some departure from the mean along the direction of PC loading, and its amount of departure is measured by the variance of principal component. In contrast to PCA, CCA does not give any clue to the directional change from mean for the associated canonical variates while their correlation is measured.

In this paper we propose a functional canonical correlation analysis that is available even for paired functions having a few measurements in irregular manner. The proposed FCCA is developed under the linear mixed-effects models, where each process is represented in smooth basis functions. Random effects are modeled in the line of canonical correlation analysis so that part of random effects become canonical variates. Therefore, processes are expressed in the linear combination of factors, some of which have canonical variates as coefficients. The factor having a canonical variate as its coefficient has a directional interpretation in the data space in the sense that it shows the functional change in one process along with the associated functional changes in the other process. In the linear mixed-effects models, processes are naturally factored into two parts: one is responsible for the association between two processes, and another accounts for the remaining process variability free from the association between two processes. This factorization leads to the natural interpretation that the whole functional variability is well separated into covariability (variability due to the association between two processes) and intra-variability (variability independent from the association). Thus, covariance function of two processes is expressed in the weighted sum of the products of canonical factors (functional components associated with canonical variates in the linear mixed-effects model) from two processes with canonical correlations as weights.

The remainder of this paper is organized as follows. In Section 2, we briefly review CCA in multivariate and functional cases and provide a new formulation of CCA for both cases. Its interpretational advantage is also discussed. FCCA model is provided in Section 3 as the linear mixed-effects model that is able to cope with irregularly and sparsely observed functional data. We discuss the maximum likelihood estimation and model selection procedure in Section 4. To demonstrate the performance of the proposed approach, the finite-sample simulation studies are conducted in Section 5. The proposed method is applied to two real datasets and their results are given in Section 6. Section 7 concludes this paper with the summary and some remarks.

2. Preliminary

2.1. Canonical correlation analysis

Suppose that $\mathbf{x}$ and $\mathbf{y}$ are p and q-dimensional random vectors with $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}_{\mathbf{x}}$, $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}_{\mathbf{y}}$, $\mathrm{Var}(\mathbf{x}) = \boldsymbol{\Sigma}_{\mathbf{x}}$, $\mathrm{Var}(\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{y}}$, and $\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{xy}}$. In classical multivariate analysis, CCA finds linear combinations $\mathbf{a}^\top\mathbf{x}$ and $\mathbf{b}^\top\mathbf{y}$ having the largest possible correlation with one another. Such $\mathbf{a} \in \mathbb{R}^p$ and $\mathbf{b} \in \mathbb{R}^q$ are found to maximize

$$\mathrm{Cov}^2\left(\mathbf{a}^\top\mathbf{x}, \mathbf{b}^\top\mathbf{y}\right)$$

subject to $\mathrm{Var}\left(\mathbf{a}^\top\mathbf{x}\right) = \mathrm{Var}\left(\mathbf{b}^\top\mathbf{y}\right) = 1$. Provided that the covariance matrices $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}$ are nonsingular, the optimal canonical weights $\mathbf{a}$ and $\mathbf{b}$ are given by $\mathbf{a} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\mathbf{u}$ and $\mathbf{b} = \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}\mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$ are obtained from the singular value decomposition (SVD) of the matrix $\mathbf{T} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}$.

Denote by $\mathbf{u}_1, \ldots, \mathbf{u}_p$ and $\mathbf{v}_1, \ldots, \mathbf{v}_q$ the left- and right-singular vectors of $\mathbf{T}$. Then, we have

$$\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) = \sum_{\ell=1}^{p} \eta_\ell \mathbf{u}_\ell, \qquad \boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}) = \sum_{\kappa=1}^{q} \xi_\kappa \mathbf{v}_\kappa \tag{2.1}$$

with $\eta_\ell = \mathbf{u}_\ell^\top\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) = \mathbf{a}_\ell^\top(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})$ and $\xi_\kappa = \mathbf{v}_\kappa^\top\boldsymbol{\Sigma}_{\mathbf{y}}^{-1/2}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}) = \mathbf{b}_\kappa^\top(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}})$. The random variables η_ℓ and ξ_ℓ, $1 \le \ell \le \min(p, q)$, are canonical variates. Multiplying $\boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}$ and $\boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}$ to both sides of two equations in (2.1) respectively,