Journal of Statistical Planning and Inference 180 (2017) 49–68

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Statistical inference for stochastic processes: Two-sample hypothesis tests

Andrea Ghiglietti [a,*], Francesca Ieva [a], Anna Maria Paganoni [b]

[a] Department of Mathematics "F. Enriques", Università degli Studi di Milano, Italy

[b] MOX – Modeling and Scientific Computing, Politecnico di Milano, Italy

ARTICLE INFO

Article history:
Received 27 April 2016
Received in revised form 16 May 2016
Accepted 16 August 2016
Available online 24 August 2016

Keywords:
Functional data
Distances in L^2
Hypothesis tests
Two-sample problems

ABSTRACT

In this paper, we present inferential procedures to compare the means of two samples of functional data. The proposed tests are based on a suitable generalization of Mahalanobis distance to the Hilbert space of square integrable functions defined on a compact interval. The only conditions required concern the moments and the independence of the functional data, while the distribution of the processes generating the data is not needed to be specified. Test procedures are proposed for both the cases of known and unknown variance–covariance structures, and asymptotic properties of test statistics are deeply studied. A simulation study and a real case data analysis are also presented.

1. Introduction

In the last two decades, the statistical research has been motivated by an increasing interest in the study of high dimensional data, having a wide range of applications in biology, chemometrics, medicine, meteorology and finance, among others. In all these cases observed data may be points that belong to functions generated by a continuous time stochastic process with values in a suitable infinite dimensional Hilbert space, typically $L^2(T)$, with T compact interval of $\mathbb{R}$. Functional Data Analysis (FDA) gathers all the statistical models and tools required for the study of this kind of data characterized by a number p of features observed for each statistical unit much larger than the sample size n (see Ramsey and Silverman, 2002, 2005; Ferraty and Vieu, 2006; Horváth and Kokoszka, 2012 and references therein). Classical methodologies in FDA are concerned with the mean function and the covariance kernel of the process generating the data. The estimation and the inference on the mean function is typically realized by computing confidence bands that take into account the covariance structure of the process, see for instance see Yao et al. (2005) and Ma et al. (2012) for sparse longitudinal data, Bunea et al. (2011), Degras (2011) and Cao et al. (2012) for dense functional data.

The problem of testing the difference between means in functional framework has been faced in different areas of the statistical research. In particular (Ramsey and Silverman, 2005) proposed a pointwise t-test, Zhang et al. (2010) introduced a test based on L^2 norm for the Behrens–Fisher problem, Pini and Vantini (2016) proposed an interval testing procedure based on permutation tests, Staicu et al. (2014) developed a pseudo likelihood ratio test and Staicu et al. (2015) studied a L^2-norm based global testing procedure for multilevel structured functional data. In a nonparametric setting (Hall and Keilegom, 2007) studied the best smoothing techniques to minimize the amount of difference in the distributions generating data, due to the reconstruction from discrete sampled observations. The two-sample comparison can be viewed also as a particular case of functional analysis of variance (FANOVA) with only two groups of sampled curves (see among others

* Corresponding author.
E-mail addresses: andrea.ghiglietti@unimi.it (A. Ghiglietti), francesca.ieva@unimi.it (F. Ieva), anna.paganoni@polimi.it (A.M. Paganoni).

http://dx.doi.org/10.1016/j.jspi.2016.08.004
0378-3758/

Cuevas et al., 2004; Abramovich and Angelini, 2006; Antoniadis and Sapatinas, 2007; Zhang, 2013 and Horváth and Rice, 2015). In (Horváth et al., 2013) this problem is addressed studying the sample covariance operator for functional time series, whereas in a classification context (Galeano et al., 2015) studied a truncated version of the Mahalanobis distance for functional data that they proved to be a semi-distance in L^2. In many cases, these procedures consist of a suitable dimensional reduction of the data, which allows classical multivariate procedures to be applied. Although this approach may be satisfactory in some situations, the functional nature of the data is not fully exploited and some information may be lost due to the dimensional reduction. Our aim is to construct a testing procedure based on a distance in L^2 that either considers all the components of data, without truncation, and takes into account the structure of the variance–covariance operator.

In Ghiglietti and Paganoni (2014), the authors proposed a procedure to test the difference between the means of Gaussian processes based on a generalization of Mahalanobis distance (say d_p) that achieves two main goals: (i) to consider all the infinite components of data basis expansion and (ii) to share the same ideas which the Mahalanobis distance is based on. In this paper, we extend those inferential procedures to a wide range of situations, by relaxing some strong assumptions made in Ghiglietti and Paganoni (2014). Specifically, the proposed tests do not require to specify the distribution of the processes generating the data and allow comparisons between samples with different covariance functions. In particular, in this work we relax the strong assumption on the Gaussianity of the processes generating data required in Ghiglietti and Paganoni (2014). In this wider context, we prove theoretical results on the convergence of the distance d_p between the sample mean and a fixed function $m(t) \in L^2(T)$ and between means of two independent samples. Additionally, we establish the rate of convergence and the exact asymptotic distributions of the distance d_p between functional sample means. The rate of convergence and the limiting distributions are sharply different when the true means of the processes are equal or different. Indeed, in the first case, we show that the exact rate of convergence is n^{-1} and the limiting distribution is a strictly positive random variable; in the second case, we prove a central limit theorem (CLT) for the distance d_p between functional sample means with Gaussian asymptotic distribution and the rate of convergence of $n^{-\frac{1}{2}}$.

The almost sure convergence established for the distance d_p of the sample means guarantees the consistency of this estimator, while the second-order asymptotic results provide the probabilistic basis to construct test procedures for the comparison of the means of two functional populations. Indeed, the proposed critical regions are based on the limiting distribution established in the case of equality of the means, while the CLT is applied to compute the asymptotic power of the test given any difference between the means of the processes. Test procedures are proposed for both the cases of known and unknown variance–covariance structures. It is worth noting that all the results hold also for multivariate functional data case.

Moreover, the functional framework adopted in the paper includes as a special case the multivariate finite dimensional setting. In particular, since in this latter case the standard Mahalanobis distance is well defined and it can be derived by a particular choice of d_p, the asymptotic results and the inferential procedures of this paper can be applied to the multivariate data analysis based on the Mahalanobis distance.

The rest of the paper is structured as follows: firstly, the distance d_p is introduced and its main properties are discussed in Section 2. Then, the asymptotic results on the behavior of this metric applied to random processes are presented in Section 3. Section 4 is concerned with inferential procedures for the comparison of the means in functional data analysis. Specifically, critical regions based on the distance d_p applied to functional sample means are proposed. Finally, a simulation study together with a real case data analysis are presented in Section 5. Appendices A and B gather the proofs of the theorems stated in Section 3. All the analyzes are carried out with R (R Development Core Team, 2009).

2. Properties of a generalized Mahalanobis distance for functional objects

In this section, we present the generalized Mahalanobis distance for infinite dimensional spaces introduced in Ghiglietti and Paganoni (2014). This metric is characterized by features similar to the Mahalanobis distance, which is commonly used in the multivariate context. Hence, inferential tools for infinite dimensional objects can be constructed in an analogous way to the procedures typically used for multivariate elements (see Section 4). We start by describing the motivation problem that required the introduction of the new distance.

Let y and w be realizations of a stochastic process $X \in L^2(T)$, where T is a compact interval of $\mathbb{R}$. Let $m(t) = \mathbb{E}[X(t)]$ be the mean function and V the covariance operator of X, i.e. V is a linear compact integral operator from $L^2(T)$ to $L^2(T)$ acting as follows: $(Va)(s) = \int_T v(s,t)a(t)dt \ \forall a \in L^2(T)$, where v is the covariance function defined as $v(s,t) = \mathbb{E}[(X(t)-m(t))(X(s)-m(s))]$. Then, denote by $\{\lambda_k; k \geq 1\}$ and $\{\varphi_k; k \geq 1\}$ the sequences of eigenvalues and eigenfunctions, respectively, associated to v.

Letting $\langle a, b\rangle = \int_T a(t)b(t)dt$ be the usual inner product in $L^2(T)$, the natural generalization of the Mahalanobis distance in the functional framework would be the following

$$d_M(y, w) = \sqrt{\sum_{k=1}^{\infty} \frac{(\langle y - w, \varphi_k\rangle)^2}{\lambda_k}}. \tag{1}$$

This distance takes into account the correlations and the variability described by the covariance structure of X. However, it is well known that in an infinite dimensional space d_M is not a proper distance in $L^2(T)$, since the series in (1) can diverge