Journal of Statistical Planning and Inference 143 (2013) 1500–1511

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

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

# Testing for a change in covariance operator

Daniela Jarušková *

*Czech Technical University, Prague, Czech Republic*

ARTICLE INFO

*Article history:*
Received 26 November 2012
Received in revised form
15 March 2013
Accepted 30 April 2013
Available online 9 May 2013

*Keywords:*
Functional data
Covariance operator
Principal components
Two-sample problem
Change point problem

ABSTRACT

The paper considers a problem of equality of two covariance operators. Using functional principal component analysis, a method for testing equality of $K$ largest eigenvalues and the corresponding eigenfunctions, together with its generalization to a corresponding change point problem is suggested. Asymptotic distributions of the test statistics are presented.



## 1. Introduction

The statistical inference where the objects of interest are random functions has recently become popular, see e.g. Bosq (2000), Ferraty (2011) and Horváth and Kokoszka (2012), because objects of interest are often continuous functions. Even if the measurements are taken discretely in many points it is assumed that the observed values are some noisy measurements of a quantity that is a continuous function in reality.

The problem that motivated our study came from civil engineering where the behavior of a tunnel primary lining thickness along the tunnel was studied. In any of $N$ profiles in distances $\{j\Delta,\ j=1,\ldots,N\}$ from the tunnel entrance the thickness was measured in $p (p \gg N)$ equidistant points from left to right. There are two ways we can proceed. As the distance $\Delta$ was relatively large we can suppose that our observations $\{\boldsymbol{X}_j = (X_j(1),\ldots,X_j(p))^T,\ j=1,\ldots,N\}$ form a sequence of independent random vectors. The other possibility is to approximate the multidimensional vectors by continuous smooth functions. Then, the observations $\{X_j(t), 0\le t\le 1\},\ j=1,\ldots,N$ form a sequence of independent random functions. The basic question may be whether the basic stochastic characteristics, i.e., the mean functions $m_j(t)=EX_j(t)$ as well as covariance operators $\mathcal{A}_j$ defined by covariance functions $A_j(s,t)=EX_j(t)X_j(s)$ remain the same for all $j=1,\ldots,N$ or whether there exists a point $j_0$ (a change point) such that the characteristics before and after the change point differ. The problem belongs to the change point analysis. The methods for detecting a change in mean function were studied by Aue, Gabrys et al. (2009). Here, we consider a problem of detecting a change in a covariance function supposing that $m_1(t)=\cdots=m_N(t)$ for $t\in[0,1]$.

Besides application in civil engineering, the problem of detecting change(s) in the stochastic behavior of functional data or random vectors with many components may be encountered in many other fields, e.g. in climatology and hydrology when stationarity of annual cycles is studied. Here, changes may be caused by all types of human activity. Heat islands of big cities not only increase mean winter temperature but also decrease its variability. Deforestation of certain areas may be a

* Tel.: +420608532915.
*E-mail address:* jarus@mat.fsv.cvut.cz

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.04.011

cause of more frequent summer floods. Many authors, e.g. Aue, Hörmann et al. (2009) or Wied et al. (2011, 2012) present applications of change point detection of covariance structures in an analysis of time series of stock indices. Horváth et al. (2010) present an application to detect breaks in a time series of credit card transactions and Galeano and Peña (2009) in price indices.

We start with a two-sample decision problem on the equality of two covariance operators $\mathcal{A}$ and $\mathcal{B}$. The problem was introduced in Benko et al. (2009) and studied by Panaretos et al. (2010). Panaretos et al. (2010) mentioned that the extension of finite dimensional procedures can lead to complications, as the infinite-dimension version of the problem constitutes an ill-posed inverse problem. This is also true when the operators $\mathcal{A}$ and $\mathcal{B}$ are finite-dimensional but the numbers of observations are smaller than their dimension. Panaretos et al. (2010) suggested to choose a set of functions $\phi_1,\dots,\phi_p$ and check whether $\langle\phi_i,(\mathcal{A}-\mathcal{B})\phi_{i'}\rangle=0$ for $1\le i\le p$, $1\le i'\le p$. It is clear that using a finite set of test functions, one cannot generally find all departures from $\mathcal{A}=\mathcal{B}$ if $\mathcal{A}$ and $\mathcal{B}$ are infinite dimensional. Even if their dimension is finite but very large, one would have to use a correspondingly large set of $\{\phi_i\}$. On the other hand, the test functions $\{\phi_i\}$ may be chosen to detect departures from $\mathcal{A}=\mathcal{B}$ that are of special interest to us. Clearly, in test procedures the functions $\{\phi_i\}$ may be chosen to be functions of observations.

In our paper we decide to check whether $\langle u_k,(\mathcal{A}-\mathcal{B})u_k\rangle=0$ as well as $\langle v_k,(\mathcal{A}-\mathcal{B})v_k\rangle=0$ for $k=1,\dots,K$, where $\{u_k\}$ are eigenfunctions of $\mathcal{A}$ that correspond to $K$ largest eigenvalues $\{\lambda_k\}$, $k=1,\dots,K$ of $\mathcal{A}$, and $\{v_k\}$ are eigenfunctions of $\mathcal{B}$ that correspond to $K$ largest eigenvalues $\{\mu_k\}$, $k=1,\dots,K$ of $\mathcal{B}$. This is true if and only if $\mathcal{A}_K=\mathcal{B}_K$, where the operator $\mathcal{A}_K$ corresponds to the function $A_K=\sum_{k=1}^{K}\lambda_k u_k(t)u_k(s)$ and the operator $\mathcal{B}_K$ corresponds to the function $B_K=\sum_{k=1}^{K}\mu_k v_k(t)v_k(s)$ and this is true if and only if $\lambda_1=\mu_1,\dots,\lambda_K=\mu_K$ and $u_1=v_1,\dots,u_K=v_K$.

It may happen that we are interested in detection of $\mathcal{A}_K\neq\mathcal{B}_K$, i.e., we consider to test the null hypothesis $\mathcal{A}_K=\mathcal{B}_K$ against the alternative $\mathcal{A}_K\neq\mathcal{B}_K$. If the hypothesis $\mathcal{A}_K=\mathcal{B}_K$ is rejected, we may be interested in the question which sources caused the rejection. If the hypothesis $\mathcal{A}_K\neq\mathcal{B}_K$ is not rejected and if $\|\mathcal{A}-\mathcal{A}_K\|$ as well as $\|\mathcal{B}-\mathcal{B}_K\|$ are relatively small, we may conclude that if $\mathcal{A}$ and $\mathcal{B}$ differ from each other, then they differ only slightly.

From many examples of principal component analysis we know that for some data a small number $K$ of sources exist that are able to express a large proportion of total variability. The favorable situation occurs when $K$ is known apriori from a similar type of data. For instance, analyzing covariance matrices of 365-dimensional vectors that correspond to smooth annual cycles of 16 small Czech rivers in the period 1935–1996, we have seen that the four largest principle components explained 76–83% and the five largest principle components 82–88% of the total variance. (The vectors were obtained by smoothing daily values by a kernel smoothing technique using the Epanechnikov window with a bandwidth of $h=15$.) These principle components explained the most important sources of variance, i.e., the time and length of the spring high discharge period caused by snow melting, and variability of mean winter discharges. If, for example, we would like to test the stability of covariance matrices of smoothed annual cycles for a small Czech river, we would use $K=5$. If we do not have such prior information we may try to choose $K$ based on proportions of $K$ principle components in the estimated total variability.

In change point analysis we usually test a null hypothesis $H_{cp}^0$ claiming that all observations have the same distribution against an alternative claiming that at some unknown time point a specific characteristic of distribution has changed. Derivation of a test statistic in change point detection usually has two steps. First, an appropriate test statistic for two-sample problems for a fixed and known change point is suggested. If the change point is unknown, one can calculate such a test statistic for any possible change point $1\le j\le N$ so that a sequence of test statistics is obtained. Then, the test statistic for an unknown change point is a certain functional of that sequence, usually a sum or a maximum. In our paper we recommend applying a sum of weighted two-sample statistics with weights that decrease with $N_1N_2/N^2$ where $N_1$ and $N_2$ corresponds to a number of observations in the first part, respectively of the second part of the series.

In the two-sample problem the asymptotic distribution under $H_0$ of the suggested test statistics is a $\chi^2$ distribution. The proof can be obtained similarly as in Panaretos et al. (2010) for Gaussian processes or in Fremdt et al. (2012) for non-Gaussian processes. In the corresponding change point problem we show that under the null hypothesis the test statistic converges in distribution to a integral of a sum of squares of independent Brownian bridges.

The paper is organized as follows. In Section 2 we suggest the test statistics for the two-sample problem in case of Gaussian processes as well as non-Gaussian processes and derive their asymptotic distribution. In Section 3 we suggest the test statistics for the corresponding change-point problem for both Gaussian processes and non-Gaussian processes and derive their asymptotic distribution. In Section 4 we present two applications. The first one comes from civil engineering and was motivation for our study of the corresponding change point problem. The second application comes from climatology and the goal of a statistic inference was a comparison of annual cycles of Milan and Padua temperature series.

## 2. Two-sample test

We observe two independent sequences of i.i.d. zero mean processes $X_1(t),\dots,X_{N_1}(t)$ and $Y_1(t),\dots,Y_{N_2}(t)$ defined for $t\in[0,1]$ such that $E\int_0^1 X_1^4(t)\,dt<\infty$ and $E\int_0^1 Y_1^4(t)\,dt<\infty$. Let $N=N_1+N_2$. We suppose that the covariance functions $A(t,s)=E\,X_1(t)X_1(s)$ and $B(t,s)=E\,Y_1(t)Y_1(s)$ are continuous functions on $[0,1]^2$. We denote the corresponding covariance operator of $X_1$ defined by the kernel $A(t,s)$ by $\mathcal{A}$ and the covariance operator of $Y_1$ defined by the kernel $B(t,s)$ by $\mathcal{B}$

$$(\mathcal{A}v)(t)=\int_0^1 A(t,s)v(s)\,ds,\quad (\mathcal{B})v(t)=\int_0^1 B(t,s)v(s)\,ds,\quad v\in L^2[0,1].$$