Journal of Statistical Planning and Inference 162 (2015) 20–42

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Curve registration by nonparametric goodness-of-fit testing

Olivier Collier, Arnak S. Dalalyan *

ENSAE ParisTech/CREST/GENES, 3 avenue P. Larousse, 92245 Malakoff, France

ARTICLE INFO

Article history:
Received 26 August 2014
Received in revised form 5 December 2014
Accepted 10 February 2015
Available online 19 February 2015

Keywords:
Nonparametric inference
Hypotheses testing
Wilks' phenomenon
Keypoint matching

ABSTRACT

The problem of curve registration appears in many different areas of applications ranging from neuroscience to road traffic modeling. In the present work,[1] we propose a nonparametric testing framework in which we develop a generalized likelihood ratio test to perform curve registration. We first prove that, under the null hypothesis, the resulting test statistic is asymptotically distributed as a chi-squared random variable. This result, often referred to as Wilks' phenomenon, provides a natural threshold for the test of a prescribed asymptotic significance level and a natural measure of lack-of-fit in terms of the p-value of the χ^2-test. We also prove that the proposed test is consistent, *i.e.*, its power is asymptotically equal to 1. Finite sample properties of the proposed methodology are demonstrated by numerical simulations. As an application, a new local descriptor for digital images is introduced and an experimental evaluation of its discriminative power is conducted.

Introduction

Boosted by applications in different areas such as biology, medicine, computer vision and road traffic forecasting, the problem of curve registration and, more particularly, some aspects of this problem related to nonparametric and semiparametric estimation, have been explored in a number of recent statistical studies. In this context, the model used for deriving statistical inference represents the input data as a finite collection of noisy signals such that each input signal is obtained from a given signal, termed mean template or structural pattern, by a parametric deformation and by adding a white noise. Hereafter, we refer to this as the *deformed mean template* model. The main difficulties for developing statistical inference in this problem are caused by the nonlinearity of the deformations and the fact that not only the deformations but also the mean template used to generate the observed data are unknown.

While the problems of estimating the mean template and the deformations were thoroughly investigated in recent years, the question of the adequacy of modeling the available data by the deformed mean template model received little attention. By the present work, we intend to fill this gap by introducing a nonparametric goodness-of-fit testing framework that allows us to propose a measure of appropriateness of a deformed mean template model. To this end, we focus our attention on the case where the only allowed deformations are translations and propose a measure of goodness-of-fit based on the p-value of a chi-squared test.

Model description

We consider the case of functional data, that is each observation is a function on a fixed interval, taken for simplicity equal to [0, 1]. More precisely, assume that two independent samples, denoted $\{X_i\}_{i=1,\ldots,n}$ and $\{X_i^{\#}\}_{i=1,\ldots,n^{\#}}$, of functional

* Corresponding author.
E-mail address: dalalyan@imagine.enpc.fr (A.S. Dalalyan).

[1] This paper was presented in part at the AI-STATS 2012 conference.

http://dx.doi.org/10.1016/j.jspi.2015.02.004
0378-3758/© 2015 Elsevier B.V. All rights reserved.

data are available such that within each sample the observations are independent identically distributed (i.i.d.) drifted and scaled Brownian motions. Let f and $f^{\#}$ be the corresponding drift functions: $f(t) = d\mathbf{E}[X_1(t)]/dt$ and $f^{\#}(t) = d\mathbf{E}[X_1^{\#}(t)]/dt$. Then, for $t \in [0, 1]$,

$$X_i(t) = \int_0^t f(u)\,du + sB_i(t), \qquad X_\ell^{\#}(t) = \int_0^t f^{\#}(u)\,du + s^{\#}B_\ell^{\#}(t),$$

where $s, s^{\#} > 0$ are the scaling parameters and $(B_1, \ldots, B_n, B_1^{\#}, \ldots, B_{n^{\#}}^{\#})$ are independent Brownian motions. Since we assume that the entire paths are observed, the scale parameters s and $s^{\#}$ can be recovered with arbitrarily small error using the quadratic variation. So, in what follows, these parameters are assumed to be known (an extension to the setting of unknown noise level is briefly discussed in Section 3).

The goal of the present work is to provide a statistical testing procedure for deciding whether the curves of the functions f and $f^{\#}$ coincide up to a translation. Considering periodic extensions of f and $f^{\#}$ on the whole real line, this is equivalent to checking the null hypothesis

$$\mathbf{H_0}: \qquad \exists\,(\tau^*, a^*) \in [0, 1] \times \mathbb{R} \quad \text{such that } f(\cdot) = f^{\#}(\cdot + \tau^*) + a^*. \tag{1}$$

If the null hypothesis is satisfied, we are in the set-up of a deformed mean template model, where $f(\cdot)$ plays the role of the mean template and spatial translations represent the set of possible deformations.

Starting from Golubev (1988) and Kneip and Gasser (1992), semiparametric and nonparametric estimation in different instances of the deformed mean template model have been intensively investigated (Rønn, 2001; Dalalyan et al., 2006; Gamboa et al., 2007; Dalalyan, 2007; Castillo and Loubes, 2009; Bigot and Gadat, 2010; Trigano et al., 2011; Castillo, 2012; Härdle and Marron, 1990; Carroll and Hall, 1992; Vimond, 2010; Castillo, 2007; Bigot et al., 2009b) with applications to image warping (Glasbey and Mardia, 2001; Bigot et al., 2009a). However, prior to estimating the common template, the deformations or any other object involved in a deformed mean template model, it is natural to check its appropriateness, which is the purpose of this work.

To achieve this goal, we first note that the pair of sequences of complex-valued random variables $\boldsymbol{Y} = (Y_0, Y_1, \ldots)$ and $\boldsymbol{Y}^{\#} = (Y_0^{\#}, Y_1^{\#}, \ldots)$, defined by

$$\big[Y_j, Y_j^{\#}\big] = \int_0^1 e^{2\pi \mathrm{i} jt}\, d\Bigg[\frac{1}{n}\sum_{i=1}^{n} X_i(t),\ \frac{1}{n^{\#}}\sum_{\ell=1}^{n^{\#}} X_\ell^{\#}(t)\Bigg],$$

is a sufficient statistic in the model generated by observations $(X_1, \ldots, X_n)$ and $(X_1^{\#}, \ldots, X_{n^{\#}}^{\#})$. Therefore, without any loss of information, the initial (functional) data can be replaced by the transformed data $(\boldsymbol{Y}, \boldsymbol{Y}^{\#})$. Let us denote by $c_j = \int_0^1 f(x)\, e^{2\mathrm{i}j\pi x}\,dx$ and $c_j^{\#} = \int_0^1 f^{\#}(x)\, e^{2\mathrm{i}j\pi x}\,dx$ the complex Fourier coefficients of the signals f and $f^{\#}$. Then, the first components of the observed sequences, $(Y_0, Y_0^{\#})$, can be written as

$$Y_0 = c_0 + \frac{s}{\sqrt{n}}\,\epsilon_0, \qquad Y_0^{\#} = c_0^{\#} + \frac{s^{\#}}{\sqrt{n^{\#}}}\,\epsilon_0^{\#},$$

where ϵ_0 and $\epsilon_0^{\#}$ are two independent, real, standard Gaussian variables. Furthermore, for $j \geq 1$, we have

$$Y_j = c_j + \frac{s}{\sqrt{2n}}\,\epsilon_j, \qquad Y_j^{\#} = c_j^{\#} + \frac{s^{\#}}{\sqrt{2n^{\#}}}\,\epsilon_j^{\#}, \tag{2}$$

where the complex valued random variables $\epsilon_j, \epsilon_j^{\#}$ are i.i.d. standard Gaussian: $\epsilon_j, \epsilon_j^{\#} \sim \mathcal{N}_{\mathbb{C}}(0, 1)$, which means that their real and imaginary parts are independent $\mathcal{N}(0, 1)$ random variables. Moreover, $(\epsilon_0, \epsilon_0^{\#})$ are independent of $\{(\epsilon_j, \epsilon_j^{\#}) : j \geq 1\}$. In what follows, we will use boldface letters for denoting vectors or infinite sequences so that, for example, $\boldsymbol{c}$ and $\boldsymbol{c}^{\#}$ refer to $\{c_j; j = 0, 1, \ldots\}$ and $\{c_j^{\#}; j = 0, 1, \ldots\}$, respectively.

Under the mild assumption that f and $f^{\#}$ are squared integrable, the likelihood ratio of the Gaussian process $\boldsymbol{Y}^{\bullet,\#} = (\boldsymbol{Y}, \boldsymbol{Y}^{\#})$ is well defined. Using the notation $\boldsymbol{c}^{\bullet,\#} = (\boldsymbol{c}, \boldsymbol{c}^{\#})$, $\sigma = s/\sqrt{2n}$ and $\sigma^{\#} = s^{\#}/\sqrt{2n^{\#}}$, the corresponding negative log-likelihood is given by

$$\ell(\boldsymbol{Y}^{\bullet,\#}, \boldsymbol{c}^{\bullet,\#}) = \frac{(Y_0 - c_0)^2}{4\sigma^2} + \frac{(Y_0^{\#} - c_0^{\#})^2}{4\sigma^{\#2}} + \sum_{j\geq 1}\bigg(\frac{|Y_j - c_j|^2}{2\sigma^2} + \frac{|Y_j^{\#} - c_j^{\#}|^2}{2\sigma^{\#2}}\bigg). \tag{3}$$

In the present work, we present a theoretical analysis of the penalized likelihood ratio test in the asymptotics of large samples, *i.e.*, when both n and $n^{\#}$ tend to infinity, or equivalently, when σ and $\sigma^{\#}$ tend to zero. The finite sample properties are examined through numerical simulations.