Journal of Statistical Planning and Inference 147 (2014) 53–65

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Predicting the continuation of a function with applications to call center data

Y. Goldberg [a,*], Y. Ritov [b], A. Mandelbaum [c]

[a] Department of Statistics, University of Haifa, Haifa 31705, Israel
[b] Department of Statistics and the Center for the Study of Rationality, The Hebrew University, Jerusalem 91905, Israel
[c] Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 24 January 2013
Received in revised form
17 November 2013
Accepted 18 November 2013
Available online 1 December 2013

Keywords:
Functional data analysis
Call center data
Workload process
B-spline
Knot insertion algorithm

ABSTRACT

We show how to construct the best linear unbiased predictor (BLUP) for the continuation of a curve, and apply the proposed estimator to real-world call center data. Using the BLUP, we demonstrate prediction of the workload process, both directly and based on prediction of the arrival counts. The Matlab code and all data sets in the presented examples are available in the supplementary material.

1. Introduction

Many data sets consist of a finite number of observations, where each of these observations is a sequence of points. It is often natural to assume that each sequence is a set of noisy measurements of points on a smooth curve. In such cases, it can be advantageous to address the observations as functional data rather than as a multiple series of data points. This approach was found useful, for example, in noise reduction, missing data handling, and in producing robust estimations (see the books of Ramsay and Silverman, 2002, 2005, for a comprehensive treatment of functional data analysis). In this work we consider the problem of forecasting the continuation of a curve using functional data techniques.

The problem we consider here is relevant to longitudinal data sets, in which each observation consists of a series of measurements over time that are sampled from an underlying curve, possibly with noise. Examples of such curves are growth curves of different individuals and arrival rates of calls to a call center or of patients to an emergency room during different days. We assume that such curves, or measurement series that approximate these curves, were collected previously. We would like to estimate the continuation of a new curve given its beginning, using the behavior of the previously collected curves.

The forecasting of curve continuation suggested here is based on finding the best linear unbiased predictor (BLUP) (Robinson, 1991). We assume that the curves are governed by a small number of underlying functional patterns, possibly with additional noise. These underlying functional patterns determine the main variation between the different curves.

* Corresponding author. Tel.: +972 48288507.
E-mail addresses: yair.goldy@gmail.com, ygoldberg@stat.haifa.ac.il (Y. Goldberg), yaacov.ritov@huji.ac.il (Y. Ritov), avim@tx.technion.ac.il (A. Mandelbaum).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.11.006

The computation of the predictor is performed in two steps. First, the underlying functional patterns' coefficients are estimated from the beginning of the new curve, which is defined on the "past" segment. Second, the prediction is obtained by computing the representation of the curve patterns on the "future" segment. We prove that the resulting estimator is indeed the BLUP and that it is a smooth continuation of the beginning of the curve (at least in the absence of noise). From a computational point of view, we discuss the use of B-splines in the representation of the curves on the different segments. We explain why a B-spline representation ensures an efficient and stable way to compute the mean function and covariance operators on different partial segments.

We apply the proposed forecasting procedure to call center data. We forecast the continuation of two processes: the arrival process and the workload process (i.e., the amount of work in the system; see, for example, Aldor-Noiman et al., 2009). In call centers, the forecast of the arrival process plays an important role in determining staffing levels. Optimization of the latter is important since salaries account for about 60–70% of the cost of running a call center (Gans et al., 2003). Usually, call center managers utilize forecasts of the arrival process and knowledge of service times, along with some understanding of customer patience characteristics (Zeltyn, 2005), to estimate future workload and determine the staffing level (Aldor-Noiman et al., 2009). The disadvantage of this approach is that the forecast of the workload is not performed directly, and instead it is obtained using the forecast of the arrival process. Reich (in preparation) showed how the workload process can be estimated explicitly, thereby enabling direct forecast of the workload. In this work, we forecast the continuation of both the arrival and workload processes, given past days' information and the information up to a certain time of day. Since the actual processes are not smooth, we first approximate these processes with smooth curves. We compare direct and indirect forecasting results for the workload process. We also compare our results for the arrival process to those of other forecasting techniques, namely, to the techniques that were introduced by Weinberg et al. (2007) and Shen and Huang (2008).

This paper has two main contributions. First, we present a novel functional-data prediction method for continuation of a curve. We show that the proposed method is the best unbiased linear predictor for continuation of a curve. The proposed estimator is fast and easy to compute, and is a continuous process in time, thus enabling prediction for any given future time. Second, we demonstrate how to predict the workload process directly, and compare this direct method to the usual indirect ones that are based on prediction of the arrival process.

Forecasting of the continuation of a function was considered in previous works. Aguilera et al. (1997) proposed to predict the continuation of the curve by regression of the principal components of the second part of the interval on the principal components of the first part of the interval. Shen (2009), in the context of time series data, proposed to first forecast the new curve entirely, and then update this forecast based on the given curve beginning. Both of these methods do not discuss curve continuity at the point dividing the interval, or optimality of the estimator. In a different context, Yao et al. (2005) proposed a functional data method for sparse longitudinal data that enables prediction of curves, even if only a few measurements are available for each curve. Although this method can be used to forecast the continuation of a curve, it was not designed to optimize such prediction. This is also evident in the case study in Section 4, in which we compare the method of Yao et al. (2005) to the proposed BLUP.

The paper is organized as follows. The functional model and notation are presented in Section 2. In Section 3 we show how to construct the BLUP for the continuation of a curve. In Section 4 we apply the estimator to real-world data, comparing direct and indirect workload forecasting, and our results to other techniques. Concluding remarks appear in Section 5. Proofs are provided in the Appendix. A link for the code and data sets used for the case study appear in the Supplemental Materials.

2. The functional framework

In this section we present the functional model and notation that will be used for the construction of the BLUP.

2.1. The functional model

Assume that we observe random i.i.d. functions $Y^{(1)},\dots,Y^{(M)}$ that are defined on the segment $S=[0,T]$. We assume that these functions have a basis expansion with respect to some N-dimensional continuous function space $\mathcal{S}$ (N can possibly be large). For now we do not impose any specific structure on the space $\mathcal{S}$, but in Sections 2.3 and 2.4 we will focus on spline functional spaces as an important example. Given a new function $Y^{(M+1)}$ which is observed only on the segment $S_1=[0,U]$, for some $0<U<T$, we would like to estimate the continuation of this function on the segment $S_2=[U,T]$.

We assume that, up to some noise, the functions $\{Y^{(m)}(t)\}_m$ are contained in some low-dimensional subspace of $\mathcal{S}$. More specifically, we assume that each function can be written as

$$Y^{(m)}(t)=\mu(t)+\sum_{i=1}^{p}h_i^{(m)}\phi_i(t)+\varepsilon^{(m)}(t)=\mu(t)+\boldsymbol{h}^{(m)'}\boldsymbol{\phi}(t)+\varepsilon^{(m)}(t), \tag{1}$$

where $\mu(t)\in\mathcal{S}$ is the mean function, $\boldsymbol{h}^{(m)}=(h_1^{(m)},\dots,h_p^{(m)})'$ is a random vector with mean zero and covariance matrix L, $\boldsymbol{\phi}(t)=(\phi_1(t),\dots,\phi_p(t))'$ is a vector of orthonormal functions in $\mathcal{S}$; $\varepsilon^{(m)}(t)$ is the noise which is defined to be the part of Y that is not in the span of the $\boldsymbol{\phi}(t)$. We assume that p, the dimension of the subspace, is much smaller than N, the dimension of $\mathcal{S}$. Such decomposition can arise, for example, when using principal component analysis for functional data (Ramsay and