Computational Statistics and Data Analysis 93 (2016) 76–85

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixtures of spatial spline regressions for clustering and classification

Hien D. Nguyen, Geoffrey J. McLachlan *, Ian A. Wood

School of Mathematics and Physics, University of Queensland, Australia

ARTICLE INFO

Article history:
Received 24 April 2013
Received in revised form 14 January 2014
Accepted 17 January 2014
Available online 25 January 2014

Keywords:
Functional data
Mixture model
Classification
Clustering
Spatial spline

ABSTRACT

Classification and clustering of functional data arise in many areas of modern research. Currently, techniques for performing such tasks have concentrated on applications to univariate functions. Such techniques can be extended to the domain of classifying and clustering bivariate functions (i.e. surfaces) over rectangular domains. This is achieved by combining the current techniques in spatial spline regression (SSR) with finite mixture models and mixed-effects models. As a result, three novel techniques have been developed: spatial spline mixed models (SSMM) for fitting populations of surfaces, mixtures of SSR (MSSR) for clustering surfaces, and MSSR discriminant analysis (MSSRDA) for classification of surfaces. Through simulations and applications to problems in handwritten character recognition, it is shown that SSMM, MSSR, and MSSRDA are effective in performing their desired tasks. It is also shown that in the context of handwritten character recognition, MSSR and MSSRDA are comparable to established methods, and are able to outperform competing approaches in missing-data situations.

1. Introduction

In recent years, functional data analysis (FDA) (Ramsay and Silverman, 1997) has become a popular tool for statistical analysis and pattern recognition. The popularity of FDA techniques has largely come from their ability to drastically reduce the dimensionality of curve-type data. Because of this fact, FDA has lent itself to numerous modern applications in the areas of biology, economics, medicine, machine learning, and sociology. Examples of such applications can be found in Ramsay and Silverman (2002).

Of particular interest in this article are the applications of FDA to problems of clustering and classifying functional data. Here, to dispel confusion, we define classification as the learning of decision rules based on labeled data from different populations, and we define clustering as the partitioning of a single population into separate subpopulations.

There is currently a large literature on techniques for clustering and classifying functional data. Interesting developments in the area include B-spline regression for classification by linear discriminant analysis (James and Hastie, 2001) and for clustering by mixtures of linear mixed models (MLMMs) (James and Sugar, 2003), Fourier basis regression for clustering by MLMMs (Ng et al., 2006), piecewise polynomial regression for clustering (Chamroukhi et al., 2010) and for classification Chamroukhi et al. (2013), Gaussian process regression for classification by principal component analysis (Hall et al., 2001) and by centroid-based methods (Delaigle and Hall, 2012), support vector machines (SVMs) for classification (Rossi and Villa, 2006), and nonparametric density estimation for clustering (Boulle, 2012).

Upon review of the literature, it is evident that the concentration in the area is towards the analysis of univariate functions. Although not as prolific, the analysis of surfaces generated from bivariate functions is also a well-developed area. Some important works in this direction are: Kriging (Matheron, 1963), thin-plate splines (Duchon, 1977), multivariate adaptive regression splines (Friedman, 1991), soap film smoothing (Wood et al., 2008), and spatial spline regression (SSR)

* Correspondence to: Department of Mathematics, University of Queensland, St. Lucia, 4072, Australia. Tel.: +61 7 3365 2150; fax: +61 7 3365 1477.
E-mail addresses: h.nguyen7@uq.edu.au (H.D. Nguyen), g.mclachlan@uq.edu.au, gjm@maths.uq.edu.au (G.J. McLachlan).

0167-9473/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2014.01.011

(Malfait and Ramsay, 2003; Ramsay et al., 2011; Sangalli et al., 2013). The developments in bivariate FDA appear focused on the aspects of smoothing and estimating data from a single surface population rather than performing inference regarding data from a population of functions.

In this article we aim to bridge the gap by extending the functional clustering technique of James and Sugar (2003) and the classification technique of James and Hastie (2001) to the case of bivariate functional data. This is done through a novel application of bivariate spline functions from Malfait and Ramsay (2003) to the problem of probability density estimation for populations of bivariate functions.

The approach that we developed uses a linear mixed-effects model (LMM) framework for modeling the distributions of bivariate functions. When finite mixtures of such distributions are constructed, the approach naturally leads to a MLMM clustering technique, similar to those of James and Sugar (2003), Celeux et al. (2005), and Ng et al. (2006). We refer to the technique described as mixtures of spatial spline regressions (MSSR).

In the cases where we have multiple populations of surfaces, we can apply the MSSR method to model each population. These population distributions can then be used to produce discrimination rules of the mixture discriminant analysis (MDA) type (Hastie and Tibshirani, 1996). We name this approach: mixtures of spatial spline regressions discriminant analysis (MSSRDA), and apply both it and MSSR to the problems of classifying and clustering handwritten characters.

We shall proceed as follows. A brief overview of the SSR approach will be given in Section 2. We describe the spatial spline mixed model framework (SSMM) which extends SSR to model populations of surfaces in Section 3. The MSSR extension of the SSMM framework is presented in Section 4, and its application to clustering and classification is described in Section 5. Simulated examples of the SSMM, MSSR, and MSSRDA techniques are given in Section 6, and an application to problems in handwritten character recognition is presented in Section 7. Finally, conclusions are drawn in Section 8.

2. Spatial spline regressions

Spatial spline regression (SSR) is a technique first introduced in Malfait and Ramsay (2003) and later elaborated upon in Ramsay et al. (2011) and Sangalli et al. (2013). In this article, we concentrate on the application of SSR in a rectangular domain as applied in the original article. The method is outlined as follows.

Let $Y_1, \ldots, Y_m$ be an i.i.d. random sample with realizations $y_1, \ldots, y_m$ such that given the coordinates $\boldsymbol{x}_k = (x_{1k}, x_{2k})^T$, we have the relationship $Y_k = \mu(\boldsymbol{x}_k) + E_k$, where $E_k \sim N(0, \sigma^2)$ and $k = 1, \ldots, m$. Here, the superscript T indicates matrix transposition, and μ is an unknown function which maps from $R = \left[x_1^-, x_1^+\right] \times \left[x_2^-, x_2^+\right]$ to $\mathbb{R}$.

If μ has a known parametric form, then techniques from nonlinear regression may be used to estimate the function from data (see Bates and Watts (1988) for details). However, we are only concerned with the case where μ is unknown in this article.

2.1. Nodal basis functions

In the univariate literature, a popular approximation for μ in a bounded domain is through the use of B-splines (de Boor, 1978). Specific details of such use can be found in Ramsay and Silverman (1997).

The idea of B-splines can be extended to the domain of surface approximation through applications of nodal basis functions (NBFs) from the finite elements literature (see for example Braess (2001)). In this article we use the linear "tent shaped" NBFs of Malfait and Ramsay (2003), which are sufficient for the tasks of clustering and classification. Higher order spatial splines for applications where smoothness is important are discussed in Sangalli et al. (2013).

The linear NBF is a function s with parameters $\boldsymbol{c} = (c_1, c_2)^T$ (center), δ_1 (horizontal shape parameter), and δ_2 (vertical shape parameter). For completeness, the exact form of the linear NBF is

$$
\begin{aligned}
&s(\boldsymbol{x}; \boldsymbol{c}, \delta_1, \delta_2) \\
&= \begin{cases}
-\dfrac{x_2}{\delta_2} + \dfrac{c_2\delta_2}{\delta_2} & \text{if } \boldsymbol{x} \in \left\{(x_1, x_2) : c_1 < x_1 \le c_1 + \delta_1, \dfrac{\delta_2}{\delta_1}x_1 + \dfrac{\delta_1 c_2 - \delta_2 c_1}{\delta_1} \le x_2 \le c_2 + \delta_2\right\}, \\
-\dfrac{x_1}{\delta_1} + \dfrac{c_1 + \delta_1}{\delta_1} & \text{if } \boldsymbol{x} \in \left\{(x_1, x_2) : c_1 < x_1 \le c_1 + \delta_1, c_2 \le x_2 < \dfrac{\delta_2}{\delta_1}x + \dfrac{\delta_1 c_2 - \delta_2 c_1}{\delta_1}\right\}, \\
-\dfrac{x_1}{\delta_1} + \dfrac{x_2}{\delta_2} + \dfrac{\delta_1\delta_2 + \delta_2 c_1 - \delta_1 c_2}{\delta_1\delta_2} & \text{if } \boldsymbol{x} \in \left\{(x_1, x_2) : c_1 \le x_1 \le c_1 + \delta_1, \dfrac{\delta_2}{\delta_1}x_1 + \dfrac{\delta_1 c_2 - \delta_2 c_1 - \delta_1\delta_2}{\delta_1} \le x_2 < c_2\right\}, \\
\dfrac{x_2}{\delta_2} + \dfrac{\delta_2 - c_2}{\delta_2} & \text{if } \boldsymbol{x} \in \left\{(x_1, x_2) : c_1 - \delta_1 \le x_1 < c_1, c_2 - \delta_2 \le, x_2 \le \dfrac{\delta_2}{\delta_1}x_1 + \dfrac{\delta_1 c_2 - \delta_2 c_1}{\delta_1}\right\}, \\
\dfrac{x_1}{\delta_1} + \dfrac{\delta_1 - c_1}{\delta_1} & \text{if } \boldsymbol{x} \in \left\{(x_1, x_2) : c_1 - \delta_1 \le x_1 < c_1, \dfrac{\delta_2}{\delta_1}x_1 + \dfrac{\delta_1 c_2 - \delta_2 c_1}{\delta_1} < x_2 \le c_2\right\}, \\
\dfrac{x_1}{\delta_1} - \dfrac{x_2}{\delta_2} + \dfrac{\delta_1\delta_2 + \delta_1 c_2 - \delta_2 c_1}{\delta_1\delta_2} & \text{if } \boldsymbol{x} \in \left\{(x_1, x_2) : c_1 - \delta_1 \le x_1 \le c_1, c_2 < x_2 \le \dfrac{\delta_2}{\delta_1}x_1 + \dfrac{\delta_1 c_2 + \delta_1\delta_2 - \delta_2 c_1}{\delta_1}\right\}, \\
0 & \text{otherwise,}
\end{cases}
\end{aligned}
\tag{1}
$$

and an example of one appears in Fig. 1.