Computational Statistics and Data Analysis 78 (2014) 56–68

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian nonparametric classification for spectroscopy data

Luis Gutiérrez[a,*], Eduardo Gutiérrez-Peña[b], Ramsés H. Mena[b]

[a] Escuela de Salud Pública, Fac. de Medicina, U. de Chile, Chile
[b] IIMAS-UNAM, Mexico

ARTICLE INFO

Article history:
Received 9 August 2013
Received in revised form 12 April 2014
Accepted 12 April 2014
Available online 19 April 2014

Keywords:
Discriminant analysis
Food authentication
Gaussian process
Geometric weights prior

ABSTRACT

High-dimensional spectroscopy data are increasingly common in many fields of science. Building classification models in this context is challenging, due not only to high dimensionality but also to high autocorrelations. A two-stage classification strategy is proposed. First, in a data pre-processing step, the dimensionality of the data is reduced using one of two distinct methods. The output of either of these methods is then used to feed a classification procedure that uses a multivariate density estimate from a Bayesian nonparametric mixture model for discrimination purposes. The model employed is based on a random probability measure with decreasing weights. This nonparametric prior is chosen so as to ease the identifiability and label switching problems inherent to these models. This simple and flexible classification strategy is applied to the well-known 'meat' data set. The results are similar or better than previously reported in the literature for the same data.

1. Introduction

Spectroscopy data are increasingly common in many fields of science. Spectroscopy is the study of the interaction between radiation and matter as a function of wavelength. It measures the reflectance or absorbance values mainly in the visible and near-infrared region of the electromagnetic spectrum. The values of reflectance are produced by vibrations in the chemical bonds in the substance analyzed. Of particular interest in this work are the spectroscopies corresponding to different types of meats. Fig. 1 shows $p = 1050$ wavelength points, reporting reflectance measurements for 110 samples of homogenized meat: $n_1 = 55$ samples of chicken (gray lines) and $n_2 = 55$ samples of turkey (green lines). Reflectance measurements were taken in the range 400–2498 nm at 2 nm intervals. These data were first reported and analyzed by McElhinney et al. (1999) with the purpose of classification between different types of meats in the context of food authentication. Authentication is the process by which food or beverages are verified to match their label description (Winterhalter, 2007).

Indeed, statistical methods for spectroscopy data are appealing in a variety of fields. See, for example, the recent statistical and computational methods for spectroscopy data in Lee and Cox (2010) and Chakraborty (2012) in the context of smoothing spectra and multiple response kernel regression with spectroscopy predictors, respectively. Spectroscopy is the source of information in many biomedical and pharmaceutical research such as cardiovascular radiology, brain imaging, quality/process control and clinical trials, among others. In particular, in the context of food authentication, discriminant analysis methods constitute a key tool (Brown et al., 1999; Dean et al., 2006; Toher et al., 2007; Gutiérrez et al., 2011). However, as is evident from Fig. 1, these data feature high autocorrelation and non-linearity, hence requiring a robust model able to capture such behavior (Murphy et al., 2010).

[*] Corresponding author.
E-mail address: luisgutierrez@med.uchile.cl (L. Gutiérrez).

http://dx.doi.org/10.1016/j.csda.2014.04.010

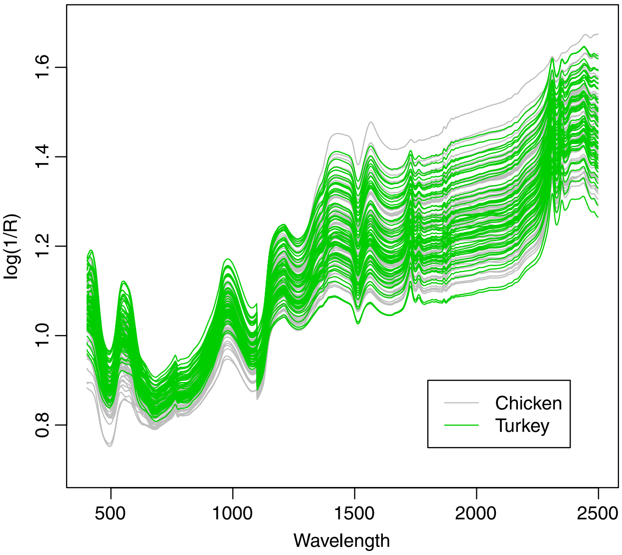

Fig. 1. Spectra curves. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In a nutshell, high dimensional spectroscopic data are characterized by: (1) the number of samples n is typically much smaller than the number of variables or measurements p ($n \ll p$); (2) the curve trajectories are non-linear; (3) the data are evenly spaced as a function of wavelength, usually every 2 nanometers; (4) the data exhibit high positive autocorrelation; (5) due to this autocorrelation, the curve trajectories tend to be smooth.

There are two common approaches to deal with the high dimensionality of data in classification contexts. The first one is to use a dimension reduction technique such as principal component analysis (Jollife, 1986), and then use only the first p^* components to assign units to groups by means of a classification method for $p^* < n$. The second approach is to first select a subset of p^* variables, useful for discrimination, and then perform the classification with a method for $p^* < n$. In both approaches, the classification method for $p^* < n$ commonly used is linear or quadratic discriminant analysis; see, for example, Fearn et al. (2002); Datta (2008); Murphy et al. (2010) and Stingo et al. (2012). However, linear and quadratic discriminant methods are based on the assumption of normality and hence are not always robust.

Motivated by the above data set, generated for food authentication purposes, a Bayesian nonparametric classification approach for spectroscopy data is proposed. The proposal is based on a two-stage procedure: first, a dimension reduction of the spectroscopy curves, from $p \gg n$ to $p^* < n$ dimensions, is conducted. Second, with the selected p^* variables, a robust version of quadratic discriminant analysis is used. In the first stage, two approaches for dimension reduction are explored: principal component analysis and a simple variable selection algorithm. The latter is based on a Gaussian process with random effects which jointly models the spectra in order to identify wavelength regions that are informative about the group membership.

In stage two, a flexible discrimination procedure based on a multivariate Bayesian nonparametric mixture model with geometric weights is used (Fuentes-García et al., 2010; Mena et al., 2011; Mena, 2013). An appealing feature of this nonparametric prior is its ordered weights construction, which leads to a more adequate distribution of the latent allocation variables reducing also some identifiability issues (Mena and Walker, 2013). Furthermore, as discussed in Fuentes-García et al. (2010), such a feature results in better posterior density estimates and therefore in better discrimination results.

Indeed, for one dimensional density estimation problems, this model has proved to be more efficient than stick-breaking based mixture models such as the Dirichlet process mixture (DPM) model (Fuentes-García et al., 2010). Hence, as a by-product of the present analysis, the efficiency of this model in multivariate settings is validated, which to the best of our knowledge has not been done elsewhere.

When this mixture is based on a Gaussian kernel, the proposed classification model can be seen as a robust generalization of Quadratic Discriminant Analysis (QDA) since it allows for asymmetries and multimodality in the distribution of the responses. Supervised classification based on mixtures of Dirichlet process has been discussed by De la Cruz-Mesía et al. (2007) in a biomedical context, and by Gutiérrez and Quintana (2011) in food authentication. The present proposal has the advantage that it is considerably simpler than other approaches based on finite mixtures or DPMs, and at the same time does not compromise any of the appealing nonparametric features.

This work is organized as follows. In Section 2, a general Bayesian classification approach is presented. This section also develops the classification equations and decision rules employed in the sections that follow. In Section 3, the dimension reduction strategy is described. Section 4 discusses the nonparametric prior distribution employed and also contextualizes it within the flexible multivariate classification model. Section 5 presents an example with a real spectroscopy data set in the context of food authentication. Finally, Section 6 contains some concluding remarks.