Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Compositional regression with functional response ☆

R. Talská [a], A. Menafoglio [b,*], J. Machalová [a], K. Hron [a], E. Fišerová [a]

[a] *Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, CZ-77146 Olomouc, Czech Republic*

[b] *MOX, Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy*

HIGHLIGHTS

- A new function-on-scalar regression model for density responses is proposed.
- The problem is faced by embedding density data in a Bayes space.
- Novel computational methods based on a B-spline representation for PDFs are discussed.
- The methodology is tested via simulation and applied to real data.

ARTICLE INFO

Article history:
Received 12 April 2017
Received in revised form 25 January 2018
Accepted 30 January 2018
Available online xxxx

Keywords:
Bayes spaces
Regression analysis
Density functions
B-spline representation

ABSTRACT

The problem of performing functional linear regression when the response variable is represented as a probability density function (PDF) is addressed. PDFs are interpreted as functional compositions, which are objects carrying primarily relative information. In this context, the unit integral constraint allows to single out one of the possible representations of a class of equivalent measures. On these bases, a function-on-scalar regression model with distributional response is proposed, by relying on the theory of Bayes Hilbert spaces. The geometry of Bayes spaces allows capturing all the key inherent features of distributional data (e.g., scale invariance, relative scale). A B-spline basis expansion combined with a functional version of the centered log-ratio transformation is utilized for actual computations. For this purpose, a new key result is proved to characterize B-spline representations in Bayes spaces. The potential of the methodological developments is shown on simulated data and a real case study, dealing with metabolomics data. A bootstrap-based study is performed for the uncertainty quantification of the obtained estimates.

1. Introduction

Distributional data in their discrete form frequently occur in many real-world surveys. For instance, frequencies of occurrence of observations from a continuous random variable – aggregated according to a given partition of the domain of observation – are typically represented by a histogram, which in turn approximates an underlying (continuous) probability density function (PDF). In general, PDFs are Borel measurable functions that are constrained to be non-negative and to integrate to unity. One may think at the unit-integral constraint as a way to single out a proper representation of the underlying measure rather than an inherent feature of PDFs themselves. Indeed, when changing the value to which the PDF integrates, to a general positive constant c (i.e., the measure of the whole), the *relative* information carried by PDFs is

☆ The codes that implement the proposed regression methods are available online as supplementary material of the present article.

* Corresponding author.
E-mail address: alessandra.menafoglio@polimi.it (A. Menafoglio).

https://doi.org/10.1016/j.csda.2018.01.018
0167-9473/© 2018 Elsevier B.V. All rights reserved.

preserved, this property being called *scale invariance* of PDFs. Here, *relative information* is to be interpreted in terms of the contributions of Borel sets of real line to the overall measure of the support of the corresponding random variable (Hron et al., 2016). Due to the peculiar features of PDFs (e.g., the aforementioned scale invariance and additional properties such as the so-called *relative scale*) the standard L^2 space of square integrable functions appears to be inappropriate for their representation. For instance, the sum of two PDFs according to the geometrical structure of the L^2 space leads to a function that is not a PDF anymore. Even more interestingly, multiplication of a PDF by a real constant yields a scaled PDF, which carries the same relative information as the original PDF according to scale invariance. The relative nature of PDFs indicates that *ratios* between values rather than absolute values represent the relevant source of information. Accordingly, instead of absolute differences, ratios between them should be considered to measure distances and dissimilarities.

In this context, Bayes (Hilbert) spaces provide a well-defined geometrical framework to represent PDFs (van den Boogaart et al., 2010, 2014; Egozcue et al., 2006). The idea motivating the introduction of Bayes spaces was to generalize the well-known Aitchison geometry for finite-dimensional compositional data (i.e., positive observations carrying exclusively relative information, Aitchison (1986) and Pawlowsky-Glahn et al. (2015)) to the infinite-dimensional setting. In fact, any PDF can be seen as a composition with infinitely many parts.

Although the general problem of functional regression has been extensively studied in the literature on functional data analysis (FDA, e.g., Ramsay and Silverman, 2005), to the best of the authors' knowledge none of the available works propose a concise methodology for regression analysis in the presence of a distributional response. In this context, this work aims to develop a general theoretical and computational setting allowing for the estimation and uncertainty assessment in linear models with a distributional response. This is relevant from both the methodological and the application-oriented viewpoints. Indeed, having at one's disposal a statistical methodology for the regression of PDF data would enable to assess the entire distribution of the response variable, rather than few statistical moments, such as the mean and the variance. Besides, it would constitute a valuable alternative to quantile regression, with the significant advantage of (a) assessing all the distribution's quantiles jointly and (b) guaranteeing that the ordering among quantiles is preserved by the estimation procedure.

The key point of the proposed approach is to consider PDFs as elements of a Bayes space, and accordingly work with the geometry of the latter space. The centered log-ratio (clr) transformation – that allows representing the PDFs through zero-integral elements of L^2 – is then used to ease computations while using the Bayes space geometry (van den Boogaart et al., 2014; Hron et al., 2016; Menafoglio et al., 2014, 2016a, b). A B-spline representation of clr-transformed data (Machalová et al., 2016) is employed to express discretely observed PDFs as smooth functions. On these bases, effective computational procedures are proposed to perform the estimations and assess their uncertainty. The potential of the proposed method shall be demonstrated through a real case study dealing with metabolite concentrations. Further, a simulation study will be introduced to assess the sensitivity of the methodology to the parameters associated with the B-spline representation (e.g., number of knots).

The remaining part of the work is organized as follows. Section 2 recalls the basic notion of Bayes spaces as mathematical spaces for PDF data. The function-on-scalar regression model is briefly recalled in Section 3 for data in L^2. A function-on-scalar model for distributional responses in Bayes spaces is discussed in Section 4. Section 5 proposes a novel computational setting – based on a B-spline representation for PDFs in Bayes spaces – which can be employed for actual computations of the proposed estimators, while Section 6 relates our findings with previous works on compositional regression for multivariate data. Section 7 tests the performances of the method through an extensive simulation study. Section 8 illustrates the application of the methodological developments to real data on metabolites concentrations, and Section 9 finally concludes the work.

2. Probability densities as elements of Bayes spaces

As for finite-dimensional compositional data, a proper choice of the sample space for PDFs is essential. Indeed, as shown in Delicado (2011) and Hron et al. (2016), analyzing PDFs within the usual L^2 space may lead to meaningless results. Instead, the peculiarities of densities can be captured through Bayes spaces, which rely upon an appropriate Hilbert space structure to deal with the data constraints.

We consider two positive functions f and g with the same support to be equivalent if $f = c \cdot g$, for a positive constant c. Recalling the scale invariance of PDFs, this implies that densities (not necessarily unit-integral densities, i.e., PDFs) within an equivalence class provide the same relative information, or, equivalently, which contributions of Borel sets to the whole mass measure do not change. For a density f, we denote by $\mathcal{C}(f)$ the unit-integral representative within its equivalence class, also named *closure*. The Bayes space $\mathcal{B}^2(I)$ consists of (equivalence classes of) densities f on a domain I for which the logarithm is square-integrable. Although the theory of van den Boogaart et al. (2014) is general and allows dealing with unbounded supports I, its construction for non-compact supports relies on reference measures different from the Lebesgue one. The latter general case raises foundational issues – both methodological and practical – which are still open. For the purpose of this work, the focus is here on the case of a compact support $I = [a, b] \subset \mathbf{R}$, which was demonstrated to be of broad applicability by several authors (Delicado, 2011; Hron et al., 2016; Menafoglio et al., 2014, 2016a, b).

In $\mathcal{B}^2(I)$, the counterparts of sum and multiplication by a scalar are called *perturbation* and *powering*, and are defined, for $f, g \in \mathcal{B}^2(I)$ and $c \in \mathbf{R}$, as

$$(f \oplus g)(t) = \frac{f(t)g(t)}{\int_a^b f(s)g(s)ds} = \mathcal{C}(fg)(t); \qquad (c \odot f)(t) = \frac{f^c(t)}{\int_a^b f^c(s)ds} = \mathcal{C}(f^c)(t),$$