Journal of Multivariate Analysis 144 (2016) 38–53

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Multivariate spline analysis for multiplicative models: Estimation, testing and application to climate change

Jean-Marc Azaïs [a], Aurélien Ribes [b,*]

[a] IMT, Univ. of Toulouse, Toulouse, France

[b] CNRM-GAME, Météo France - CNRS, Toulouse, France

ARTICLE INFO

Article history:
Received 9 January 2014
Available online 17 October 2015

AMS subject classifications:
62H
62P

Keywords:
Smoothing spline
Multiplicative models
Climate change

ABSTRACT

This paper presents multiplicative, or bi-additive, models with some spline-type regularity for a rectangular array of data, for example in space and time. The one-dimensional smoothing spline model is extended to this multiplicative model including regularity in each dimension. For estimation, we prove the existence of the maximum penalized likelihood estimates (MPLEs), and introduce a numerical algorithm that converges in a weak sense to a critical point of the penalized likelihood. Explicit MPLEs are given in two important particular cases.

With regard to hypothesis testing, we focus on the "no effect" test and prove that the null distribution of the penalized likelihood ratio test (PLRT) does not depend on the nuisance parameters under $\mathcal{H}_0$, leading to easy Monte-Carlo techniques. Numerical results are presented for both simulated data and climate data. For simulated data, our estimation algorithm is shown to have a good behavior. The application to climate data illustrates how multivariate spline analysis for multiplicative models may be of interest in climate change detection.

1. Introduction

The paper presents a multivariate model, along with its estimation and testing procedures, that can be applied to climate change detection.

Studying the evolution of climate change over a set of locations is an important issue for climate monitoring. In order to assess recent changes, many methods used in climate sciences rely on specified spatial, temporal, or spatio-temporal patterns of change. These are usually provided by numerical simulations performed with climate models (see, e.g., [12] for a review). In this paper, we take a different approach and propose to estimate the change, as well as its significance, only from observed data. To do so, we use a space–time separability assumption, following Mitchell [18] or Ribes et al. [20]. More precisely, we assume that the change depends on both space and time, but with the shape of the spatial pattern remaining constant over time, or equivalently, with the shape of the temporal pattern remaining constant over locations. Such an assumption has been widely used in climate science under the name of "pattern scaling" [18,23], e.g., in order to estimate transient climate change over the 21st century. In this way, we assume the climate change term to follow a multiplicative or biadditive model as in [8,9]. Such models are commonly used to describe interaction between two predictors. However, our model is somewhat more complex (see Section 2) than described in these papers, as correlations are assumed. This is

* Corresponding author.
E-mail addresses: jean-marc.azais@math.univ-toulouse.fr (J.-M. Azaïs), aurelien.ribes@meteo.fr (A. Ribes).

http://dx.doi.org/10.1016/j.jmva.2015.09.026

important for application to temperature data, as spatial dependence is strong. Although independence in time is sometimes invoked for annual mean temperature over land regions (e.g., [24]), we here assume a more general space–time separable covariance structure.

In the context of this multiplicative model, our statistical model assumes smooth response functions, and we therefore consider spline functions. Our multiplicative spline model is a multivariate generalization of classical spline models (see, e.g., [25]), but differs from the general multivariate form. In general, multivariate spline models are an extension of piecewise regression over simplexes, where continuity conditions are imposed to the derivatives on the edges. A presentation of bivariate splines on a lattice is given by Chui and Wang [4]. A very pedagogical presentation of the problem is also provided by Friedman [10], with an introduction to the Multivariate Adaptive Regression Spline (MARS) algorithm. In the forward procedure of the MARS algorithm, at each step, the domain is partitioned based on a covariable and a hinge function is introduced to separate two sub-regions. This step is followed by a backward elimination. Another option to fit smooth functions is provided by boosted trees methods, based on partition on bootstrap sub-samples [2,21]. The latter are usually considered as more efficient, but MARS provides a much faster algorithm. However, we have not followed this route because the space–time separability, i.e. the multiplicative structure, plays a key role in our application.

A first result is given in Section 2.2 regarding the non estimability of our multiplicative model without regularity assumption. To overcome this difficulty, we require regularity in time and/or space; our main result is that maximum penalized likelihood estimates exist (Theorem 1). These estimates, however, are not explicit; Section 3 ends with the presentation of a "flip-flop type" recursive estimation algorithm and a proof of its convergence in a weak sense. Explicit formulas for the maximum penalized likelihood estimators are given for two important particular cases: the classical univariate case and the case in which the variance–covariance matrix between the different variables is the identity, up to a scalar.

The "no effect" test is considered in Section 5. In the context of climate change, this test aims to determine whether the climate has changed or not. In IPCC language, such a test might be regarded as a "detection" test, as it is consistent with the IPCC's definition: "*Detection of change is [...] the process of demonstrating that climate [...] has changed in some defined statistical sense without providing a reason for that change.*" [11,13]. Recently, the "no effect" test (which assesses whether the regression function is zero), and the polynomial test (which assesses whether the regression function is a polynomial) have been studied in particular detail for spline models [15,6, and others]. Here we propose a penalized likelihood ratio test, and show that, under some conditions, its null distribution does not depend on the unknown parameters of the model. Consequently, a p-value can be computed after easily simulating the null distribution with Monte-Carlo techniques. This hypothesis testing procedure is different from those previously proposed in the univariate case, and could be of interest even in this simpler case.

Application of this method to multivariate simulated data (Section 6.1) illustrates the capabilities and the accuracy of the method. Application to real data (Section 6.2) provides an estimation of the pattern of temperature warming, and strong evidence that the climate has recently changed over France.

2. Statistical framework

2.1. Model and motivations

We consider a bi-dimensional array of data $Y_{i,j}$, $i = 1, \dots, n$, $\quad j = 1, \dots, p$, that represents, for example, observations at different times and places. A general functional multiplicative interaction model is assumed:

$$Y_{i,j} = f(x_{2,j}) + g(x_{1,i})\, h(x_{2,j}) + \varepsilon_{i,j}, \quad i = 1, \dots, n,\ j = 1, \dots, p, \tag{1}$$

where $f(\cdot)$, $g(\cdot)$, $h(\cdot)$ are three unknown functions, $0 \le x_{1,1} < \cdots < x_{1,n} \le 1$, and $0 \le x_{2,1} < \cdots < x_{2,p} \le 1$, are known real numbers that can be random or deterministic (e.g., equally spaced), and ε is a centered random noise. This general model is examined in Appendix A.2, but the novelty of this paper lies primarily in the estimation and hypothesis testing of the interaction term $g(x_{1,i})\, h(x_{2,j})$, and so for the sake of simplicity, we will consider the simpler version where

$$Y_{i,j} = g(x_{1,i})\, h(x_{2,j}) + \varepsilon_{i,j}. \tag{2}$$

It is assumed that $\varepsilon = (\varepsilon_{i,j})_{i,j}$ follows a Gaussian distribution with variance–covariance matrix satisfying

$$\operatorname{var}(\varepsilon) = \Sigma_g \otimes \Sigma_h, \quad \text{i.e.}\ \operatorname{cov}(\varepsilon_{i,j}, \varepsilon_{i',j'}) = (\Sigma_g)_{i,i'}(\Sigma_h)_{j,j'}, \tag{3}$$

where Σ_g and Σ_h are real symmetric positive definite matrices, of size $n \times n$ and $p \times p$ respectively. Such class of distributions has been described as multilinear normal distribution [19]. The problem where both Σ_g and Σ_h are unknown is difficult to handle and has very few practical applications. Indeed, because of dimension considerations, full-rank estimates of these two matrices cannot be obtained simultaneously as soon as $n \neq p$. Therefore we will assume $n \ge p$ (without loss of generality), and Σ_g to be known, while Σ_h is unknown.

Without loss of generality, we may fix a free multiplicative constant between Σ_g and Σ_h, and set

$$\operatorname{var}(\varepsilon) = \Sigma_g \otimes \Sigma_h = \sigma^2\, U_g \otimes U_h, \tag{4}$$