Computational Statistics and Data Analysis 55 (2011) 2871–2879

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric regression with shape-constrained penalized splines

Martin L. Hazelton [a,*], Berwin A. Turlach [b]

[a] Institute of Fundamental Sciences—Statistics & Bioinformatics, Massey University PN461, Private Bag 11222, Palmerston North, New Zealand

[b] School of Mathematics and Statistics (M019), University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

ARTICLE INFO

Article history:
Received 18 June 2010
Received in revised form 31 March 2011
Accepted 27 April 2011
Available online 4 May 2011

Keywords:
Linear mixed model
MCMC
Shape constraint
Spline
Truncated multivariate normal

ABSTRACT

In semiparametric regression models, penalized splines can be used to describe complex, non-linear relationships between the mean response and covariates. In some applications it is desirable to restrict the shape of the splines so as to enforce properties such as monotonicity or convexity on regression functions. We describe a method for imposing such shape constraints on penalized splines within a linear mixed model framework. We employ Markov chain Monte Carlo (MCMC) methods for model fitting, using a truncated prior distribution to impose the requisite shape restrictions. We develop a computationally efficient MCMC sampler by using a correspondingly truncated multivariate normal proposal distribution, which is a restricted version of the approximate sampling distribution of the model parameters in an unconstrained version of the model. We also describe a cheap approximation to this methodology that can be applied for shape-constrained scatterplot smoothing. Our methods are illustrated through two applications, the first involving the length of dugongs and the second concerned with growth curves for sitka spruce trees.

1. Introduction

Classical linear regression models are not appropriate whenever the application at hand includes correlated responses or the regression functions do not have a simple form (such as a low-order polynomial). One way of addressing these problems simultaneously is to use semiparametric regression models based on penalized splines in a linear mixed modelling framework. The inclusion of random effects allows the dependence structure in the data to be modelled, and provides requisite shrinkage estimators of spline coefficients. An excellent account of this type of semiparametric regression model is provided by Ruppert et al. (2003).

Penalized splines allow a great deal of flexibility in modelling the shape of the relationship between variables. However, this can lead to estimation of functions with implausible properties. Consider, for example, Fig. 1. This displays data on the growth of sitka spruce trees (*Picea sitchensis*) in normal conditions and in an ozone-rich atmosphere. A semiparametric model was fitted with the effect of ozone represented as a linear (fixed) effect, and penalized splines were used to describe the growth curves. The fitted model suggests that the trees decrease in size from about day 400 to day 500, and again after day 650. Such behaviour is biologically unlikely, indicating that this feature of the model is an artefact of the underlying fitted penalized spline. We revisit this example in greater detail in Section 4.2.

In cases like the sitka growth example, there is a need to place constraints on the shape of the regression function. A monotonicity constraint is a common requirement, but in some applications (particularly in economics) we may instead wish to enforce convexity (Hildreth, 1954; Matzkin, 1991). Other applications, apart from growth curves (Silverman, 1985; Silverman and Wood, 1987), in which one wishes to impose monotonicity constraints include nonparametric calibration (see, e.g., Knafl et al., 1984), and the estimation of dose–response curves (see, e.g., Kelly and Rice, 1990).

* Corresponding author. Tel.: +64 6 356 9099x2483; fax: +64 6 350 5682.
E-mail addresses: m.hazelton@massey.ac.nz (M.L. Hazelton), berwin@maths.uwa.edu.au (B.A. Turlach).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.04.018

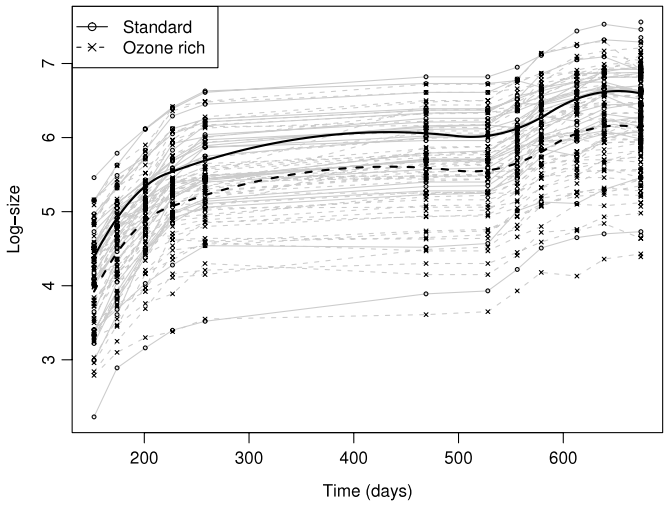

Fig. 1. A semiparametric regression model (bold lines) for the growth curves of sitka spruces (*Picea sitchensis*) in normal and ozone-rich atmospheres.

A number of authors have looked at shape-constrained spline smoothing. Some of the early notable contributions to this problem are due to Dierckx (1980), Wright and Wegman (1980), and Villalobos and Wahba (1987). Traditionally, shape-constrained spline smoothing is implemented by choosing a suitable basis for the spline, usually a *B*-spline basis, such that the desired shape constraints can easily be imposed by imposing suitable linear constraints on the parameters of the basis functions. Thus, fitting a shape-constrained smoothing spline is reduced to either a linear programming problem or a quadratic programming problem. See, among others, Micchelli et al. (1985), Irvine et al. (1986), Ramsay (1988), Fritsch (1990), Schmidt and Scholz (1990), Tantiyaswasdikul and Woodroofe (1994), and He and Shi (1998).

Alternative approaches are discussed in Ramsay (1998), Heckman and Ramsay (2000), Turlach (2005), Tutz and Leitenstorfer (2007), Leitenstorfer and Tutz (2007), Wang and Li (2008), and Wang (2008). Theoretical results can be found in Utreras (1985); Mammen and Thomas-Agnan (1999), Meyer and Woodroofe (2000), and Meyer (2008).

Later papers have often taken a Bayesian approach to the problem. See, for example, Holmes and Heard (2003), Neelon and Dunson (2004), and Shively et al. (2009). In a recent paper, Brezger and Steiner (2008) developed a method for fitting monotone functions in general additive models using Bayesian *P*-splines, building on the work of Lang and Brezger (2004) and Brezger and Lang (2006). Using a *B*-spline basis, they showed that the requirement for monotonicity is simply an isotonic ordered set of spline coefficients. They enforced this by assigning a suitably truncated multivariate normal distribution on these parameters. Model fitting was performed using Markov chain Monte Carlo (MCMC) methods. Their algorithm has a nested structure in that it is necessary to run a short inner sampler (with burn in) to draw values from the requisite truncated multivariate normal distribution, and these must be done at each step of the main (outer) MCMC algorithm.

In this paper, we also develop a shape-constrained semiparametric regression model using penalized splines, implemented in a Bayesian framework. Unlike Brezger and Steiner (2008), however, we employ a truncated power series basis for the penalized splines. We discuss how constraints can be imposed, and provide specific details for monotonic and convex regression. As with Brezger and Steiner (2008), our constraints result in the need to sample spline coefficients from a truncated normal distribution. We propose a computationally efficient MCMC sampler that does not require an inner sample to handle the truncated distribution. Our trick is to use a proposal distribution derived as a truncated version of the sampling distribution for model parameters for an unconstrained model. We also describe a cheap approximation that can be applied to scatterplot smoothing problems.

We set up our modelling framework in the next section. We begin by describing the model structure. We cover model fitting using least squares and likelihood methods because we require these results for later developments. We then discuss how shape constraints may be imposed on the regression function by placing linear constraints on the spline coefficients. In Section 3, we discuss Bayesian inference and describe our MCMC sampler. Our methods are illustrated in Section 4 on two applications, one involving the length of dugongs (a large marine mammal) and the other being a continuation of our examination of the sitka growth data. We draw conclusions and discuss possible extensions of our work in Section 5.

2. A shape-constrained semiparametric regression model

2.1. Linear mixed models with penalized splines

We start by considering the simple case where we observe data (x_i, y_i) for individuals $i = 1, \ldots, n$, and wish to model the mean of y as a function of x. A nonparametric regression model is

$$Y_i = m(x_i) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{1}$$