Computational Statistics and Data Analysis 103 (2016) 316–329

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A flexible approach to inference in semiparametric regression models with correlated errors using Gaussian processes

Heping He [a], Thomas A. Severini [b,*]

[a] College of Finance and Statistics, Hunan University (North Campus), Changsha, Hunan 410082, PR China

[b] Department of Statistics, Northwestern University, 2006 Sheridan Road, Evanston, IL 60208, USA

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form 9 May 2016
Accepted 20 May 2016
Available online 6 June 2016

Keywords:
Semiparametric model
Gaussian process regression
Generalized least squares
Restricted maximum likelihood

ABSTRACT

Consider a semiparametric regression model in which the mean function depends on a finite-dimensional regression parameter as the parameter of interest and an unknown function as a nuisance parameter. A method of inference in such models is proposed, using a type of integrated likelihood in which the unknown function is eliminated by averaging with respect to a given distribution, which we take to be a Gaussian process with a covariance function chosen to reflect the assumptions about the function. This approach is easily implemented and can be applied to a wide range of models using the same basic methodology. The consistency and asymptotic normality of the estimator of the parameter of interest are established under mild conditions. The proposed method is illustrated on several examples.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Consider semiparametric regression models in which the mean of the response depends on a finite-dimensional regression parameter β as the parameter of interest as well as an unknown function $\gamma(\cdot)$ as a nuisance parameter; in addition, such models generally contain a parameter ϕ parameterizing the covariance structure of the errors in the model. Our goal is non-Bayesian inference for the parameter β in the presence of the unknown function γ.

In the usual approaches to inference in semiparametric regression models, the function γ is estimated using some type of smoothing method, such as kernel estimation or splines. See e.g., Eubank (1999), Hastie and Tibshirani (1990), Hastie et al. (2001), Ruppert et al. (2003), and Wahba (1990) for general discussions of estimation in semiparametric regression models along with many specific results.

Here we use likelihood inference for β, eliminating γ by treating it as a Gaussian process and averaging over it, forming a type of integrated likelihood (Berger et al., 1998; Severini, 2007; He and Severini, 2014). We refer to this methodology as *Gaussian process semiparametric regression* (GPSR).

It is important to emphasize that the Gaussian process assumption for γ is used only as a method of eliminating it from the likelihood. In particular, we study the proposed estimators through their frequency properties in the model in which γ is an unknown fixed function.

The GPSR approach has several advantages over more traditional methods of estimation in semiparametric models. One is that the integrated likelihood is based on observation of a Gaussian random vector, with a linear specification for the mean and a parametric covariance matrix; thus, the calculations use standard statistical software for linear models with a

* Corresponding author.
E-mail addresses: heping.he@outlook.com (H. He), severini@northwestern.edu (T.A. Severini).

http://dx.doi.org/10.1016/j.csda.2016.05.010
0167-9473/© 2016 Elsevier B.V. All rights reserved.

parametric covariance structure. Parameters appearing in the covariance function of the Gaussian process, which control the amount of smoothing, are handled automatically; a separate method to choose the "smoothing parameter" is not needed. This is particularly useful in models with correlated errors. It is well-known that choosing the smoothing parameter for semiparametric regression models is often difficult (Opsomer et al., 2001). Although some progress has been made (see, e.g., De Branter et al., 2011; Yao and Li, 2013), it is still a challenging problem in many models.

More importantly, the methodology can be applied in any model in which the integrated likelihood can be determined, making it useful in more complex models. Such a determination generally requires properties of Gaussian random functions, which are well-studied with many results available; see, for example, Ash and Gardner (1975), Laning and Battin (1956), and Rasmussen and Williams (2006). For instance, the same basic method can be used in a wide range of models, including the partially linear model (Engle et al., 1986; Heckman, 1986; Speckman, 1988; Lin and Carroll, 2001), shape-invariant models (Lawton et al., 1972; Härdle, 1990, Ch. 9), varying-coefficient models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999), and models depending on the unknown function through a linear functional (Vardi and Lee, 1993); in addition, it can handle any type of parametric correlation structure for any of these models. The goal of this paper is to show that how the Gaussian process approach can be used in many semiparametric regression models by appropriately modeling the covariance structure of the data. The proposed methodology is applied to a number of examples, illustrating the usefulness of the GPSR method.

The outline of the paper is as follows. In Section 2, a review of Gaussian process regression and its application to semiparametric regression models is given. The application of the GPSR method to specific semiparametric regression models is considered in Section 3. The asymptotic properties of the estimators of the parametric components of the model are presented in Section 4. Section 5 considers several numerical examples. Technical details are given in the Appendix.

2. Review of Gaussian process regression

2.1. Nonparametric regression

Consider the model $Y_j = \gamma(z_j) + \epsilon_j, j = 1, \ldots, n$, where γ is an unknown function, $z_1, \ldots, z_n$ are fixed constants taking values in a set $\mathcal{Z}$, and $\epsilon_1, \ldots, \epsilon_n$ are independent normal random variables each with mean 0 and standard deviation σ. Gaussian process regression is based on modeling γ a Gaussian process with mean function 0 and covariance function $K_\lambda(\cdot, \cdot)$, where, for each λ, K_λ is a real-valued function on $\mathcal{Z} \times \mathcal{Z}$ and $\lambda \in \Lambda$ is an unknown parameter vector. In addition, $\gamma(\cdot)$ and the errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are assumed to be uncorrelated. This distribution is often viewed as a type of prior distribution for γ.

Under the Gaussian process distribution, $(\gamma(z_1), \ldots, \gamma(z_n))^{\mathrm{T}}$ has a multivariate normal distribution with mean vector 0 and covariance matrix based on

$$\mathrm{Cov}(\gamma(z_j), \gamma(z_k)) = K_\lambda(z_j, z_k), \quad j, k = 1, \ldots, n. \tag{1}$$

It follows that marginal distribution of $Y \equiv (Y_1, \ldots, Y_n)^{\mathrm{T}}$ is multivariate normal with mean vector 0 and covariance matrix $\sigma^2 I_n + \Sigma_\lambda$ where I_n is the $n \times n$ identity matrix and Σ_λ is the $n \times n$ matrix with (j, k)th element given by (1).

For $z^* \in \mathcal{Z}$, the conditional expected value of $\gamma(z^*)$ given the data is given by $\Sigma_\lambda^*(\sigma^2 I_n + \Sigma_\lambda)^{-1} Y$ where Σ_λ^* is the $1 \times n$ matrix representing the covariance matrix of $\gamma(z^*)$ and $(\gamma(z_1), \ldots, \gamma(z_n))^{\mathrm{T}}$; this may be viewed as a type of Best Linear Unbiased Predictor (BLUP). This expression, as z^* varies, yields the Gaussian process regression estimator of γ, with any unknown parameters replaced by estimators. Many estimators have been proposed for these, including likelihood-based estimators as well as those based on cross-validation. See, Zhu et al. (1998), Williams (1999), Seeger (2004), Rasmussen and Williams (2006) and Murphy (2012) for general discussions of Gaussian process regression. MacKay (1999) and Sundararajan and Keerthi (2001) discuss the problem of choosing the parameters of the covariance function; Shi and Choi (2011) discuss Gaussian process regression in models for functional data.

The properties of $K_\lambda(\cdot, \cdot)$ are chosen to reflect the assumptions regarding the function $\gamma(\cdot)$. Suppose that $\mathcal{Z}$ is a subset of the real line. Then we generally assume that the covariance of $\gamma(z)$ and $\gamma(\tilde{z})$ is a decreasing function of $|z - \tilde{z}|$; specifically, take $K_\lambda(z, \tilde{z}) = \tau^2 \bar{K}(|z - \tilde{z}|/\alpha)$ where $\bar{K}$ is a decreasing, positive definite function on $[0, \infty)$ with $\bar{K}(0) = 1$ and τ and α are positive parameters, controlling the vertical and horizontal variation of the function, respectively.

The smoothness of a Gaussian process depends on the smoothness of the function $\bar{K}$: if $\bar{K}^{(2m)}(0)$ exists and is finite, then the process is m-times mean-square differentiable (Ash and Gardner, 1975). One choice for $\bar{K}$ is the Gaussian covariance function, $\bar{K}(t) = \exp(-t^2/2)$, $t \geq 0$. Since this function is infinitely-differentiable, the Gaussian process is infinitely-differentiable in mean-square. For simplicity, here the covariance function of the Gaussian process is always taken to be the Gaussian covariance function. However, other covariance functions could be used without changing the basic methodology; see Abrahamsen (1997) and Rasmussen and Williams (2006, Chapter 4) for further discussion.

It is well-known that the Gaussian process approach to nonparametric regression is related to the methods based on splines. See, e.g., Ruppert et al. (2003), Kimeldorf and Wahba (1970) and Wahba (1990). In the spline approach, the covariance function is chosen so that the estimate of the unknown function has the desired form.

Download English Version:

https://daneshyari.com/en/article/415280

Download Persian Version:

https://daneshyari.com/article/415280

Daneshyari.com