Journal of Statistical Planning and Inference 163 (2015) 1–20

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Difference-based variance estimation in nonparametric regression with repeated measurement data

Wenlin Dai[a], Yanyuan Ma[b], Tiejun Tong[a,*], Lixing Zhu[a]

[a] *Department of Mathematics, Hong Kong Baptist University, Hong Kong*

[b] *Department of Statistics, University of South Carolina, Columbia, SC 29208, USA*

ARTICLE INFO

Article history:
Received 30 May 2014
Received in revised form 31 October 2014
Accepted 25 February 2015
Available online 9 March 2015

Keywords:
Asymptotic normality
Difference-based estimator
Least squares
Nonparametric regression
Repeated measurements
Residual variance

ABSTRACT

Over the past three decades, interest in cheap yet competitive variance estimators in nonparametric regression has grown tremendously. One family of estimators which has risen to meet the task is the difference-based estimators. Unlike their residual-based counterparts, difference-based estimators do not require estimating the mean function and are therefore popular in practice. This work further develops the difference-based estimators in the repeated measurement setting for nonparametric regression models. Three difference-based methods are proposed for the variance estimation under both balanced and unbalanced repeated measurement settings: the sample variance method, the partitioning method, and the sequencing method. Both their asymptotic properties and finite sample performance are explored. The sequencing method is shown to be the most adaptive while the sample variance method and the partitioning method are shown to outperform in certain cases.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Consider the nonparametric regression model with repeated measurement data,

$$Y_{ij} = f(x_i) + \varepsilon_{ij}, \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, m, \tag{1}$$

where Y_{ij} are observations, x_i are design points, f is an unknown mean function, and ε_{ij} are independent and identically distributed (i.i.d.) random errors with mean zero and variance σ^2. In this paper we are interested in estimating the residual variance σ^2. Needless to say, an accurate estimate of σ^2 is desired in many situations, e.g., in testing the goodness of fit and in deciding the amount of smoothing (Carroll, 1987; Carroll and Ruppert, 1988; Eubank and Spiegelman, 1990; Gasser et al., 1991). Over the past three decades, interest in cheap yet competitive variance estimates in the nonparametric setting has grown tremendously. One family of estimators which has generated great interest and has become an important tool for this purpose is the difference-based estimators. Unlike their residual-based counterparts, difference-based estimators do not require the estimation of the mean function, which involves nonparametric estimation procedures, and have therefore become quite popular in practice.

In the simple situation when $m = 1$, there already exist a large body of difference-based estimators in the literature (Dette et al., 1998). In this case, model (1) reduces to

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

[*] Corresponding author.
E-mail address: tongt@hkbu.edu.hk (T. Tong).

http://dx.doi.org/10.1016/j.jspi.2015.02.010
0378-3758/© 2015 Elsevier B.V. All rights reserved.

where Y_i are observations, and ε_i are i.i.d. random errors with mean zero and variance σ^2. Assume that $0 \le x_1 \le \cdots \le x_n \le 1$, and define the order of a difference-based estimator to be the number of observations involved in calculating a local residual. von Neumann (1941) and Rice (1984) proposed the following first-order estimator,

$$\hat{\sigma}_{\mathrm{R}}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (Y_i - Y_{i-1})^2. \tag{3}$$

Gasser et al. (1986) and Hall et al. (1990) extended the idea behind the first-order estimator and proposed some higher order difference-based estimators. Dette et al. (1998) pointed out that none of the fixed order difference-based estimators can achieve the same asymptotically optimal rate as that is achieved by the residual-based estimators (Hall and Marron, 1990). Müller et al. (2003), Tong et al. (2008) and Du and Schick (2009) proposed covariate-matched U-statistic estimators for the residual variance.

Recently, Tong and Wang (2005) and Tong et al. (2013) proposed some least squares methods for estimating the residual variance, motivated by the fact that the Rice estimator (3) is always positively biased. For the equally-spaced design, let

$$\hat{\sigma}_{\mathrm{R}}^2(r) = \frac{1}{2(n-r)} \sum_{i=r+1}^{n} (Y_i - Y_{i-r})^2, \quad r = 1, 2, \ldots.$$

Assuming that f has a bounded first derivative, they showed that $E\{\hat{\sigma}_{\mathrm{R}}^2(r)\} \simeq \sigma^2 + Jd_r + o(d_r)$, where $d_r = r^2/n^2$ and $J = \int_0^1 \{f'(x)\}^2 dx/2$. To reduce the positive bias Jd_r, they constructed a linear regression model

$$\hat{\sigma}_{\mathrm{R}}^2(r) = \sigma^2 + Jd_r + \xi_r, \quad r = 1, 2, \ldots, r_0, \tag{4}$$

where ξ_r are random errors and $r_0 = o(n)$ is the chosen bandwidth. Let $N = nr_0 - r_0(r_0+1)/2$ be the total number of difference pairs involved in (4). They assigned $w_r = (n-r)/N$ as the weight of $\hat{\sigma}_{\mathrm{R}}^2(r)$, and estimated the residual variance as the intercept through the weighted least squares regression. They further showed that the asymptotic optimal bandwidth is $h_{opt} = \{28n\sigma^4/\mathrm{Var}(\varepsilon^2)\}^{1/2}$ with the corresponding mean squared error (MSE) as

$$\mathrm{MSE}(h_{opt}) = \frac{1}{n}\mathrm{Var}(\varepsilon^2) + \frac{9\sqrt{7}}{28n^{3/2}}\sigma^2\{\mathrm{Var}(\varepsilon^2)\}^{1/2} + o\left(\frac{1}{n^{3/2}}\right).$$

When $m > 1$, we have repeated measurements. Repeated measurement data are commonly available in many statistical problems. How to take advantage of the repeated measurements and develop a variance estimator that has the same advantage of not requiring a mean estimation is of great importance. Despite the rich literature on difference-based variance estimation for model (2), very little attention has been paid to model (1) with $m \ge 2$. Gasser et al. (1986) encountered the multiple measurements issue, but they decided to order the data sequentially and treat them as if they came from different design points. Thus, the multiple measurements feature is ignored. This is quite a pity, since intuitively the repeated measurement data contain different type of information, and this new information should be taken into account in constructing estimators. We suspect that one reason very few work is available for treating multiple observations in difference based variance estimation literature is that it is not easy to combine the between-design-point difference and the within-design-point difference properly. In addition, even if a certain new treatment is proposed, it is not straightforward to analyze how effective this treatment is in theory. For example, it is difficult to know if the treatment has optimal large sample property, in other words, it is difficult to know if a better method can be found in treating the multiple measurements, either within the difference based method family or overall. In this work, we will fill this literature in both aspects. Specifically, we will propose three new difference based methods to utilize the multiple measurements, respectively the sample variance method, the partitioning method and the sequencing method. We analyze these methods and illustrate the practical advantages of each method under different data structures and/or model assumptions. In addition, we will show that one of our proposals, the sequencing method is indeed optimal in that it is root-n consistent and it reaches the minimum asymptotic estimation variability among all possible consistent estimators.

The rest of the paper is organized as follows. In Section 2, we propose three difference-based methods for estimating σ^2 in nonparametric regression with repeated measurement data: the sample variance method, the partitioning method, and the sequencing method. We also explore their asymptotic properties, especially for the proposed sequencing estimator, where we derive its MSE, its optimal bandwidth and its asymptotic normality. In Section 3, we derive the optimal efficiency bound of any estimation procedure and show that the proposed sequencing estimator reaches this universal optimal efficiency bound. Extensive simulation studies are conducted in Section 4 to evaluate and compare the finite sample performance of the proposed estimators to the residual-based estimator. We then extend the methods to the nonparametric regression models with unbalanced repeated measurement data in Section 5. Also, we demonstrate the practical application of proposed methods with one real data example in Section 6. Finally, we conclude the paper in Section 7 with a brief discussion and provide all the technical proofs in the Appendices.

Download English Version:

https://daneshyari.com/en/article/1147638

Download Persian Version:

https://daneshyari.com/article/1147638

Daneshyari.com