Biometrika (2005), **92**, 4, *pp.* 821–830

Printed in Great Britain

Estimating residual variance in nonparametric regression using least squares

By TIEJUN TONG and YUEDONG WANG

Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106, U.S.A.

tong@pstat.ucsb.edu yuedong@pstat.ucsb.edu

Summary

We propose a new estimator for the error variance in a nonparametric regression model. We estimate the error variance as the intercept in a simple linear regression model with squared differences of paired observations as the dependent variable and squared distances between the paired covariates as the regressor. For the special case of a one-dimensional domain with equally spaced design points, we show that our method reaches an asymptotic optimal rate which is not achieved by some existing methods. We conduct extensive simulations to evaluate finite-sample performance of our method and compare it with existing methods. Our method can be extended to nonparametric regression models with multivariate functions defined on arbitrary subsets of normed spaces, possibly observed on unequally spaced or clustered designed points.

Some key words: Bandwidth; Difference-based estimator; Least squares; Nonparametric regression; Quadratic form; Residual variance.

1. Introduction

Consider a nonparametric regression model

$$y_i = g(x_i) + \varepsilon_i \quad (1 \leqslant i \leqslant n),$$

where the y_i's are observations, g is an unknown mean function, and the ε_i's are independent and identically distributed random errors with zero mean and variance σ^2.

Usually one fits the mean function g first and then estimates the variance σ^2 from residual sum of squares (Wahba, 1990; Müller & Stadtmüller, 1987; Hall & Carroll, 1989; Carter & Eagleson, 1992; Neumann, 1994). However, it is often desirable to have an accurate estimator of σ^2, independent of that obtained by curve fitting, for the purpose of testing the goodness of fit or choosing the amount of smoothing (Eubank & Spiegelman, 1990; Rice, 1984; Gasser et al., 1991; Kulasekera & Gallagher, 2002). An accurate estimator of σ^2 can also be used to estimate the detection limits of immunoassay (Carroll, 1987; Carroll & Ruppert, 1988).

Most estimators of σ^2 proposed in the literature are quadratic forms of the response vector $y = (y_1, \ldots, y_n)^{\mathrm{T}}$,

$$\hat{\sigma}_D^2 = y^{\mathrm{T}} D y / \mathrm{tr}(D). \quad (1)$$