Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 340 (2005) 901–904
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistics/Probability Theory

Non-parametric estimation of the average growth curve from quantized observations and correlated errors

Karim Benhenni, Mustapha Rachdi

Université de Grenoble, UFR SHS, BP. 47, 38040 Grenoble cedex 09, France

Received 11 June 2004; accepted after revision 18 April 2005

Available online 23 May 2005

Presented by Paul Deheuvels

Abstract

In this Note, we consider the problem of estimating the regression function for a fixed design model, when we only have access to quantized and correlated data. In order for the constructed estimate to be consistent, we assume that repeated observations are available. We give the asymptotic performance in terms of the mean squared error for the regression function estimator constructed from quantized observations, and we generate the optimal bandwidth. ***To cite this article: K. Benhenni, M. Rachdi, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

Résumé

Estimation de la courbe de croissance pour des observations quantifiées et des erreurs corrélées. Dans cette Note, nous considérons le problème d'estimation de la courbe de croissance pour des données quantifiées et corrélées. Afin que l'estimateur construit soit consistant, nous supposons disposer d'observations répetées. Nous donnons le comportement asymptotique de l'estimateur construit à partir de données quantifiées et nous déduisons la largeur de fenêtre optimale. ***Pour citer cet article : K. Benhenni, M. Rachdi, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

1. Introduction

Quantization of measurement is well-recognized as a source of measurement error by engineers and metrologists in the areas of communication, information theory and signal processing. However, it is typically ignored by most statisticians as they develop methods of statistical inference, whose inputs in any real application are potentially subject to quantization effects. Most standard statistical methods treat numerical data as if they were exact, real

E-mail addresses: Karim.Benhenni@upmf-grenoble.fr (K. Benhenni), mrachdi@upmf-grenoble.fr (M. Rachdi).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2005.04.035

observations. The most common form of quantization is rounding-off, which occurs in all digital systems. A general quantizer approximates an observed value by the nearest among a finite number of representative values.

It is thus important to see how the estimation in statistical problems can be affected when one has access to quantized data instead of the true data.

A survey of the theory and technics of the quantization is given in Gray and Neuhoff [4] and some applications to statistics can be found in Lee and Vardeman [7–9]. Cambanis and Gerr [2] derived the best asymptotic optimal quantizers by using the percentiles of a probability density that minimizes the asymptotic mean squared error among a class of positive density functions. Benhenni and Cambanis [1] considered the approximation of a random integral from quantized observations when the process to integrate has a covariance that behaves like a Wiener covariance along the diagonal.

The growth curve model is useful especially for growths of animals and plants and is applied extensively to biostatistics, medical research and epidemiology, and considered by many authors such as Von Rosen [12]. Piegorsch and Bailer [11] studied the estimation of the area under the growth curve, known as the concentration-time curve in pharmacokinetic research, based on the drug's concentration at different sites, repeated measures, within the organism. The non-parametric regression model with correlated errors was considered by many authors, such as Lin and Carroll [10] among others. These authors considered different modifications of kernel construction of the non-parametric regression estimator to improve the efficiency over the standard kernel estimator when correlated observations are introduced.

In this Note, we consider the statistical problem of estimating the average growth curve for a fixed design model. We consider m experimental units, each of them having n measurements of the response:

$$Y_j(x_i) = f(x_i) + \varepsilon_j(x_i) \quad \text{where } j = 1, \ldots, m \text{ and } i = 1, \ldots, n$$

where f is the unknown average growth curve and (ε_j) is the error process.

The sampling points $\{x_i,\ i = 1, \ldots, n\}$ are usually taken equally spaced in time, series data, but other type of sampling designs can also be considered such as deterministic regular (non-uniform) designs and random designs. Although repeated measurements can naturally arise in practical situations, they can make the estimators of the curve f asymptotically consistent, as was pointed out by Hart and Wehrly [6] and the comments of Härdle [5].

We estimate consistently f from the noisy observations $\{Q(Y_j(x_i))\colon\ i = 1, \ldots, n \text{ and } j = 1, \ldots, m\}$, when the errors are correlated, where Q is the quantization function and the x_i's are known constants such that $0 \leqslant x_1 < x_2 < \cdots < x_n \leqslant 1$.

The error processes $\{\varepsilon_j,\ j = 1, \ldots, m\}$ are assumed to be centered, Gaussian, uncorrelated and weakly stationary with the same autocovariance function: $\operatorname{cov}(\varepsilon_j(x), \varepsilon_l(y)) = \rho(x - y)$ if $j = l$ and 0 if $j \neq l$.

Furthermore, the autocovariance function ρ verifies a Hölder condition of order $\alpha > 0$. Then ρ can be expanded around 0 as follows:

$$\rho(t) = \begin{cases} \rho(0) - \lambda|t|^{\alpha} + \mathrm{o}\big(|t|^{\alpha}\big) & \text{for } 0 < \alpha < 2, \\ \rho(0) + \dfrac{|t|^2}{2}\rho''(0) + \mathrm{o}(t^2) & \text{for } \alpha \geqslant 2 \end{cases} \tag{1}$$

for some $\lambda > 0$ and $\rho''(0) \neq 0$.

In this work, we give the asymptotic performance in terms of the mean squared error for the average growth curve estimator constructed from quantized observations, and we generate the asymptotically optimal bandwidth which depends on the regularity of the process through the parameters α and λ, the number of replications m, and the number of levels of quantization N.

2. Estimation of the average growth curve from the quantized observations with correlated errors

The quantization system is determined by the levels of quantization $z_1 < z_2 < \cdots < z_N$ and by the bounds of the intervals $y_1 < y_2 < \cdots < y_N$. The levels of quantization are the percentiles of a continuous positive probability

density function p_Q given by $\int_{z_{k-1}}^{z_k} p_Q(t)\,\mathrm{d}t = 1/N$ for $k = 2, \ldots, N$ and the intervals of quantization are defined by $y_k = (z_{k-1} + z_k)/2$ for $k = 2, \ldots, N$.

Then, the quantization function is defined by: $Q(y) = z_k$ when $y_k < y < y_{k+1}$ where $-\infty = y_1 < z_1 < y_2 < z_2 < \cdots < y_N < z_N < y_{N+1} = +\infty$ (see, Gray and Neuhoff [4]).

We consider the estimator of f constructed from the noisy observations $\{Q(Y_j(x_i))$: $i = 1, \ldots, n$ and $j = 1, \ldots, m\}$:

$$\hat{f}_{Q,h}(x) = \frac{1}{n}\sum_{i=1}^{n} W_{h,i}(x)\overline{Z}(x_i) \quad \text{with } \overline{Z}(x) = \frac{1}{m}\sum_{j=1}^{m} Q\big(Y_j(x)\big)$$

where the weights are such that: $W_{h,i}(x) = n\int_{m_{i-1}}^{m_i} K_h(x-u)\,\mathrm{d}u$ and the midpoints $\{m_i,\ i = 0, \ldots, n\}$ are defined by $m_0 = 0$, $m_i = (x_i + x_{i+1})/2$, for $i = 1, \ldots, n-1$ and $m_n = 1$ with $K_h(x) = 1/h\,K(x/h)$. The kernel K is an even, Lipschitz function, with support $[-1, 1]$ and $\int_{-1}^{1} K(v)\,\mathrm{d}v = 1$, and $h = h(n, m)$ is the bandwidth, such that: $h \geqslant 0$ and $\lim_{n,m\to+\infty} h = 0$.

We assume that the error process (ε_j) is centered, Gaussian and weakly stationary with covariance ρ satisfying hypothesis (1). It can be shown that, the quantized process $Q(Y(x))$ has an autocovariance function ρ_Q that can be expanded around 0 as follows:

$$\rho_Q(t) = \begin{cases} \rho_Q(0) - \lambda_N|t|^{\alpha/2} + \mathrm{o}\big(|t|^{\alpha/2}\big) & \text{for } 0 < \alpha < 2, \\ \rho_Q(0) - \beta_N|t| + \mathrm{o}\big(|t|\big) & \text{for } \alpha \geqslant 2 \end{cases}$$

where $\lambda_N = (\frac{\alpha\sqrt{\lambda}}{2\sqrt{\pi\rho(0)}})B_N$ and $\beta_N = (-\frac{\rho''(0)}{2\pi\rho(0)})^{1/2}B_N$, with $B_N = \frac{1}{\sqrt{2\pi}}\sum_{k=2}^{N}(z_k - z_{k-1})^2\exp(-y_k^2/2)$.

The optimal levels of quantization z_k^*, $k = 1, \ldots, N$, correspond to the percentiles of the asymptotically optimal Gaussian density p_Q^* with mean 0 and variance 3.

The following theorem and its corollary give the asymptotic behavior of $\hat{f}_{Q,h}$ and the optimal choice of the bandwidth h when the observations are quantized according to p_Q^*.

Theorem 2.1. *We assume that the covariance function ρ of the error process satisfies hypothesis* (1) *and f is a twice differentiable continuous function on* $[0, 1]$ *with $f''(x) \neq 0$. Then, as n, m and $N \to +\infty$:*

$$\mathbb{E}\big(\hat{f}_{Q,h}(x) - f(x)\big)^2 = \frac{1}{m}\left(\rho_Q(0) - \frac{h^{\gamma/2}}{N}b(\gamma)C_K\left(\frac{\gamma}{2}\right)\right) + \frac{h^4}{4}d_K^2\big(f''(x)\big)^2 + \mathrm{O}\left(\frac{1}{Nmn^{\gamma/2}} + \frac{h^2}{N} + \frac{1}{nN} + \frac{1}{N^2}\right) + \mathrm{o}\left(h^4 + \frac{h^{\gamma/2}}{Nm}\right)$$

where $b(\gamma) = 3\alpha(\lambda/4\rho(0))^{1/2}$ for $0 < \alpha < 2$ and $b(\gamma) = 3(-\rho''(0)/2\rho(0))^{1/2}$ for $\alpha \geqslant 2$, and where $d_K = \int_{-1}^{1} u^2K(u)\,\mathrm{d}u$ and $C_K(\gamma) = \int_{-1}^{1}\int_{-1}^{1}|u-v|^{\gamma}K(u)K(v)\,\mathrm{d}u\,\mathrm{d}v$.

Corollary 2.2. *Under the hypotheses of Theorem* 2.1*, the mean squared error is asymptotically minimum for the following choice of the bandwidth*:

$$h_Q^* = \begin{cases} \left(\dfrac{3\alpha^2\sqrt{\lambda/\rho(0)}C_K(\alpha/2)}{4d_K^2(f''(x))^2}\right)^{2/(8-\alpha)}(mN)^{-2/(8-\alpha)} & \text{if } 0 < \alpha < 2, \\ \left(\dfrac{3\sqrt{-\rho''(0)/\rho(0)}C_K(1)}{d_K^2(f''(x))^2}\right)^{1/3}(mN)^{-1/3} & \text{if } \alpha \geqslant 2. \end{cases}$$

Remark 1.

(i) The no quantized estimator (see Gasser and Müller [3]) $\hat{f}_h$ is obtained by taking $Q(Y) = Y$ in the expression of $\hat{f}_{Q,h}$. We showed that the corresponding mean squared error has the following asymptotic form:

$$\mathbb{E}\big(\hat{f}_h(x) - f(x)\big)^2 = \frac{1}{m}\big(\rho(0) - a(\gamma)h^{\alpha}\gamma C_K(\gamma)\big) + \frac{h^4}{4}\big(f''(x)\big)^2 d_K^2 + \mathrm{O}\left(\frac{1}{n^{\gamma}m} + \frac{h^2}{n}\right) + \mathrm{o}\left(h^4 + \frac{h^{\gamma}}{m}\right)$$

with $\gamma = \min(\alpha, 2)$ and $a(\gamma) = \lambda$ for $0 < \alpha < 2$ and $a(\gamma) = -\rho''(0)/2$ for $\alpha \geqslant 2$. Moreover, if $m/n = \mathrm{O}(1)$, then

$$h^* = \begin{cases} \left(\dfrac{\lambda\alpha C_K(\alpha)}{d_K^2 (f''(x))^2}\right)^{1/(4-\alpha)} m^{-1/(4-\alpha)} & \text{for } 0 < \alpha < 2, \\ \left(\dfrac{-2\rho''(0)}{d_K (f''(x))^2}\right)^{1/2} m^{-1/2} & \text{for } \alpha \geqslant 2. \end{cases}$$

(ii) We studied and compared through an Uhlenbek–Ornstein error process the performance between $\hat{f}_{Q,h}$ and $\hat{f}_h$ for different growth curves. We noticed that, the estimator $\hat{f}_h$ outperforms the quantized estimator $\hat{f}_{Q,h}$, especially when the number of levels of quantization is small. However, the two estimators have about the same performance in estimating the curve f when N is large.

Notice that, results on simulations and proofs of theorems can be requested from the authors.

References

[1] K. Benhenni, S. Cambanis, The effect of quantization on the performance of sampling designs, IEEE Trans. Inform. Theory 44 (5) (1998) 1981–1992.
[2] S. Cambanis, N.L. Gerr, A simple class of asymptotically optimal quantizers, IEEE Trans. Inform. Theory IT-29 (1983) 666–676.
[3] T. Gasser, M.G. Müller, Estimating regression functions and their derivatives by the kernel method, Scand. J. Statist. 11 (1984) 171–185.
[4] R.M. Gray, D.L. Neuhoff, Quantization, IEEE Trans. Inform. Theory 44 (6) (1998) 2325–2383.
[5] W. Härdle, Applied Nonparametric Regression, vol. 19, Cambridge University Press, Cambridge, 1989.
[6] J. Hart, T. Wehrly, Kernel regression estimation using repeated measurements data, J. Amer. Statist. Assoc. 81 (1986) 1080–1088.
[7] C.S. Lee, S.B. Vardeman, Interval estimation of a normal process mean from rounded data, J. Qual. Technol. 33 (2001) 335–348.
[8] C.S. Lee, S.B. Vardeman, Likelihood-based statistical estimation from quantized data, IEEE Trans. Instrum. Meas. 54 (1) (2005) 409–414.
[9] C.S. Lee, S.B. Vardeman, Interval estimation of a normal process standard deviation from rounded data, Commun. Statist. Simulat. 31 (2002) 13–34.
[10] X. Lin, R. Carroll, Nonparametric function estimation for clustered data when the predictor is measured without/with error, J. Amer. Statist. Assoc. 95 (450) (2000) 520–534.
[11] W. Piegorsch, A. Bailer, Minimum mean-square error quadrature, J. Statist. Comput. Simulat. 46 (1993) 217–234.
[12] D. Von Rosen, The growth curve model: a review, Commun. Statist. Theory Method 20 (9) (1991) 2791–2822.