Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 340 (2005) 615–618
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistics/Probability Theory

Stepwise sampling procedure for estimating random averages

Karim Benhenni [a], Yingcai Su [b]

[a] *Université de Grenoble, UFR SHS, BP. 47, 38040 Grenoble cedex 09, France*
[b] *Southwest Missouri State University, Springfield, MO 65804, USA*

Received 27 June 2004; accepted after revision 1 March 2005

Available online 18 April 2005

Presented by Paul Deheuvels

Abstract

The aim of this Note is to present an optimal stepwise method for estimating an integral of a time series from observations at appropriately designed sampling points. Optimal linear estimators along with sampling points are constructed via a stepwise procedure. At each stage, one term is added to the existing estimator with the addition of one new sample, and previous observations and calculations are preserved. The stepwise method is also considered when simple linear nonparametric estimators are used. Asymptotically, an optimal one-step ahead sampling point is derived by maximizing an objective function that depends on the singularity of the process at the previous points. ***To cite this article: K. Benhenni, Y. Su, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

Résumé

Estimation de moyennes aléatoires par une procédure d'échantillonnage pas à pas. Dans cette Note, on considère l'estimation de l'intégrale d'un processus stochastique à partir d'observations engendrées par une procédure optimale d'échantillonnage pas à pas. À travers cette procédure on construit des estimateurs linéaires optimaux ainsi que les points d'observations. À chaque étape de la procédure, l'estimateur actuel est modifié par l'addition d'un terme engendré par le nouveau point et permet ainsi de préserver les observations et les calculs précédents. On applique aussi cette procédure d'échantillonnage pour construire des estimateurs linéaires nonparamétriques. On montre que le point d'échantillonnage optimal asymptotique de l'étape suivante de la procédure est celui qui maximise une fonction objective qui dépend de la singularité du processus à travers sa fonction d'autocovariance aux points précédents. ***Pour citer cet article : K. Benhenni, Y. Su, C. R. Acad. Sci. Paris, Ser. I 340 (2005).***

1. Introduction

The problem of interest is to estimate an integral of a time series from observations at a finite number of appropriately designed sampling points. The performance of an estimator is measured by the mean squared error.

E-mail addresses: Karim.Benhenni@upmf-grenoble.fr (K. Benhenni), yis780f@smsu.edu (Y. Su).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2005.03.003

Optimal and simple linear estimators are considered and sampling points are so chosen to minimize the mean squared error. In view of the literature, deterministic sampling schemes can be classified into single-stage methods and stepwise methods. Single-stage methods find a fixed number of sampling points simultaneously while stepwise methods select one sampling point at a time. Stepwise procedures are desirable in practice since previous observations and calculations are preserved. At each step it is only necessary to add one term to the existing estimator if one wishes to use one additional sample.

Asymptotically optimal single-stage sampling designs are considered by many authors for different statistical problems. In particular, the estimation of regression coefficients is studied in Sacks and Ylvisaker [5]. The estimation of integrals of time series is studied in Benhenni and Cambanis [2] for those with zero or higher order quadratic mean derivatives and in Stein [6] and Pitt, Robeva and Wang [4], Benhenni [1], Istas and Laredo [3] for those with derivatives up to an arbitrary order (not necessarily an integer).

The aim of this Note is to introduce an optimal stepwise method for the estimation of an integral of a time series. The problem of estimating an integral of time series by use of both an optimal linear estimator and a simple linear estimator is introduced in Section 2. Optimal linear estimators require the precise knowledge of the covariance function while simple linear estimators depend only on observations and corresponding sampling points. In Section 3, an optimal stepwise method for estimating an integral of a time series is presented. In the proposed stepwise approach, optimal linear estimators along with sampling points are determined stepwisely to minimize the mean squared error. For a time series with no quadratic mean derivative, such as time series with Wiener, Gauss–Markov covariance, it turns out that asymptotically, an optimal one-step ahead sampling point is one of the midpoints of intervals, determined by the currently existing sampling points. In Section 4, we show that asymptotically, the rule of determining optimal stepwise sampling points for simple linear nonparametric estimators is essentially similar to that for optimal linear estimators.

2. Estimation of an integral of time series

Let $X(t),\ t \in [0,1]$ be a second order time series with zero mean $EX(t) = 0$ and covariance function $EX(t)X(s) = R(s,t)$. Consider the integral $I(X) = \int_0^1 X(t)\,\mathrm{d}t$ of X over the bounded interval $[0,1]$. Define a function f on $[0,1]$ by $f(t) = EX(t)I(X) = \int_0^1 R(s,t)\,\mathrm{d}s,\ t \in [0,1]$ and denote its integration on $[0,1]$ by $\sigma^2 = \int_0^1 f(t)\,\mathrm{d}t = \int_0^1\int_0^1 R(s,t)\,\mathrm{d}s\,\mathrm{d}t = EI^2(X)$. The function $f(t)$ plays an essential role in our discussion and the quantity σ^2 represents the total variation of the integral $I(X)$.

Here, we want to estimate the integral $I(X)$ from observations of X at n sample points $T_n = \{t_i\}_{i=1}^n \subset [0,1]$ using a linear estimator $L_n(X) = \sum_{i=1}^n c_i X(t_i) = C_n' X_n$, where $X_n' = (X(t_1), \ldots, X(t_n))$ are observations of X at T_n and $C_n' = (c_1, \ldots, c_n)$ are coefficients to be selected. The mean squared error (MSE) of the estimator $L_n(X)$ is

$$MSE(C_n/T_n) = E\big(L_n(X) - I(X)\big)^2 = \sigma^2 - 2C_n' f_n + C_n' R_n C_n,$$

where $f_n' = (f(t_1), \ldots, f(t_n))$ are the values of f at T_n and $R_n = (R(t_i,t_j))_{n\times n}$ is the variance–covariance matrix of X_n. The inverse of R_n is assumed to exist for every n. It is desired to choose the coefficients C_n and the sampling points T_n in such a way that the resulting MSE is as close to zero as possible.

For a fixed sampling design T_n, optimal coefficients $\widehat{C}_n$ minimize $MSE(C_n/T_n)$ over all possible coefficients, and are $\widehat{C}_n' = f_n' R_n^{-1}$. The optimal linear estimator and its MSE are

$$\hat{L}_n(X) = f_n' R_n^{-1} X_n, \qquad MSE(\widehat{C}_n/T_n) = \sigma^2 - f_n' R_n^{-1} f_n,$$

respectively. Clearly, for $\hat{L}_n(X)$, a sampling design T_n that maximizes $f_n' R_n^{-1} f_n$ minimizes $MSE(\widehat{C}_n, T_n)$.

The optimal linear estimator requires the complete knowledge of covariance R. To bypass this constraint, we construct a simple linear estimator that depends only on the observations at the sampling points. Without loss of generality, assume $0 \leqslant t_1 < \cdots < t_n \leqslant 1$. By applying the trapezoidal rule for integral approximation in each of

the intervals (t_i, t_{i+1}), $i = 1, \ldots, n-1$ and using the rectangular rule in the two end intervals $[0, t_1)$ and $(t_n, 1]$, a simple linear estimator is obtained as follows

$$\bar{L}_n(X) = t_1 X(t_1) + \sum_{i=1}^{n-1} (t_{i+1} - t_i)\big[X(t_i) + X(t_{i+1})\big]/2 + (1 - t_n)X(t_n).$$

The simple linear estimator $\bar{L}_n(X)$ is of the form: $\bar{L}_n(X) = \overline{C}'_n X_n$ where $\overline{C}'_n = (\bar{c}_1, \ldots, \bar{c}_n)$ with $\bar{c}_1 = (t_1 + t_2)/2$, $\bar{c}_i = (t_{i+1} - t_{i-1})/2$, $i = 2, \ldots, n-1$ and $\bar{c}_n = (2 - t_{n-1} - t_n)/2$.

3. Optimal stepwise sampling designs for optimal linear estimators

The stepwise sampling method selects one sampling point at a time. The observation at the $(n+1)$th sampling point explains the largest portion of variation of $I(X)$ unexplained by observations at the preceding n sampling points. This process continues until no significant improvement on the explained variation from new points occurs.

In general, if the first n sampling points $t_1^o, \ldots, t_n^o$ have been selected, the $(n+1)$th sampling point t_{n+1}^o is so chosen that the marginal increase of variation due to the observation $X(t_{n+1}^o)$ is maximized among all $X(t)$, $t \notin \{t_i^o\}_{i=1}^n$. The marginal increase of variation due to the observation $X(t_{n+1})$ of X, given that $X(t_1^o), \ldots, X(t_n^o)$ are already obtained, is

$$c^2\big(t_{n+1} \mid T_n^o\big) = \big[f(t_{n+1}) - f_n' R_n^{-1} r_n(t_{n+1})\big]^2 \big/ \big[R(t_{n+1}, t_{n+1}) - r_n'(t_{n+1}) R_n^{-1} r_n(t_{n+1})\big],$$

where $f_n' = (f(t_1^o), \ldots, f(t_n^o))$, $R_n = (R(t_i^o, t_j^o))_{n\times n}$ and $r_n'(t_{n+1}) = (R(t_1, t_{n+1}), \ldots, R(t_n, t_{n+1}))$. Then t_{n+1}^o is a maximizer of $c^2(t_{n+1} \mid T_n^o)$, namely,

$$c^2\big(t_{n+1}^o \mid T_n^o\big) = \sup_{t_{n+1} \notin T_n^o} c^2\big(t_{n+1} \mid T_n^o\big).$$

The corresponding mean squared error of $\hat{L}(X \mid T_{n+1}^o)$ is

$$MSE\big(\widehat{C}_n \mid T_{n+1}^o\big) = \sigma^2 - E\hat{L}^2\big(X \mid T_n^o\big) = \sigma^2 - \big[c^2(t_1^o) + c^2\big(t_2^o \mid t_1^o\big) + \cdots + c^2\big(t_{n+1}^o \mid T_n^o\big)\big].$$

Consider the class of time series with no quadratic mean derivative. The covariance $R(u, v)$, however, satisfies the following assumption:

Assumption 1. $R(u, v)$ is assumed to have continuous mixed partial derivatives up to order two off the diagonal $u \neq v$ in the unit square, and continuous limits for its first order derivative at the diagonal $u = v$ from above and below, denoted by $R^{(0,1)}(u, u\pm) = \lim_{v \to u\pm 0} \partial R(u, v)/\partial v$. The jump function of $R^{(0,1)}$ along the diagonal, $\alpha(u) = R^{(0,1)}(u, u-) - R^{(0,1)}(u, u+)$, is assumed to be continuous and not identical to zero. In addition, we require $R^{(0,2)}(u, \cdot)$ to belong to the reproducing kernel Hilbert space spanned by R for every $0 \leqslant u \leqslant 1$.

For the Wiener covariance $R(s, t) = \min(s, t)$, the jump function $\alpha(t) \equiv 1$ and $R^{(0,2)}(t, \cdot) \equiv 0$. For the Gauss–Markov covariance $R(s, t) = \exp(-|s - t|)$, $\alpha(t) \equiv 2$ and $R^{(0,2)}(t, \cdot) = R(t, \cdot)$. A class of covariance functions with nonconstant jump functions is easy to give.

Specifically, suppose that $T_n = \{t_i\}_{i=1}^n$ in $[0, 1]$ is a currently operating set of sampling points. Denote the ordered t_i by $0 \leqslant s_1 < \cdots < s_n \leqslant 1$. The precise stepwise method finds t_{n+1}^o by maximizing the exact $c^2(t \mid T_n)$ in $[0, 1]$. Here we find a $(n+1)$th sampling point by maximizing asymptotic expressions of $c^2(t \mid T_n)$ as stated in the following result.

Theorem 3.1. *Consider the problem of estimating $I(X)$ by the optimal linear estimator $\hat{L}(X \mid T_{n+1})$, where t_{n+1} is a sampling point at one-step ahead. For $t, u \in [0, 1]$, $t \neq u$, let*

$$c^2(t \mid u) = \big[f(t) - R(t, u) f(u)/R(u, u)\big]^2 \big/ \big[R(t, t) - R^2(t, u)/R(u, u)\big].$$

Denote the ordered t_i by s_i, then under Assumption 1, *an asymptotic optimal sampling point at one-step ahead, denoted by t^a_{n+1}, is a maximizer of $c^2(t \mid s_1)$ in $t \in [0, s_1)$, or one of the midpoints*: $(s_k + s_{k+1})/2$, $k = 1, \ldots, n-1$, *or a maximizer of $c^2(t \mid s_n)$ in $(s_n, 1]$; t^a_{n+1} corresponds to the largest value among the local maximum of $c^2(t \mid s_1)$ in $[0, s_1)$, $\alpha(s_k)(s_{k+1} - s_k)^3/16$, $k = 1, \ldots, n-1$ and the local maximum of $c^2(t \mid s_n)$ in $(s_n, 1]$. In addition*

$$\lim_{n\to\infty} c^2(t^a_{n+1} \mid T_n)/c^2(t^o_{n+1} \mid T_n) = 1.$$

Theorem 3.1 says that asymptotically, an optimal sampling point t^a_{n+1} at next step is one of the midpoints of intervals determined by $\{t_i\}_{i=1}^n$ and moreover t^a_{n+1} has the same performance as an exact optimal point t^o_{n+1}.

4. Optimal stepwise sampling designs for simple nonparametric linear estimators

Given a set of previously determined sampling points $T_n = \{t_i\}_{i=1}^n$, we try to find the next optimal point t_{n+1} when the simple linear estimator $\bar{L}_n(X)$ is used for estimating the integral $I(X)$.

Denote the ordered sample points T_n by $0 \leqslant s_1 < \cdots < s_n \leqslant 1$ and write $d_i = s_{i+1} - s_i$, $i = 0, \ldots, n$ with $s_0 = 0, s_{n+1} = 1$. The simple nonparametric linear estimator $\bar{L}(X \mid T_{n+1})$ is constructed stepwisely according to whether $t_{n+1} \in [0, s_1)$, (s_k, s_{k+1}), $k = 1, \ldots, n-1$, or $(s_n, 1]$.

The following result gives asymptotically the optimal stepwise sampling points for the class of time series with no quadratic mean derivatives when simple linear estimators are used. The corresponding asymptotic MSE is also obtained.

Theorem 4.1. *Consider the problem of estimating $I(X)$ by the simple linear estimator $\bar{L}(X \mid T_{n+1})$. Then under Assumption* 1, *an asymptotic optimal one-step forward sampling point t^a_{n+1} is one of the points*: $s_1/3$ *in* $[0, s_1)$, $(s_k + s_{k+1})/2$ *in* $(s_k, s_{k+1}), k = 1, \ldots, n-1$ *and* $(s_n + 2)/3$ *in* $(s_n, 1]$. *It corresponds to the largest value among* $(2s_1/3)^3\alpha(s_1) \equiv q_0$, $(s_{k+1} - s_k)^3\alpha(s_k)/16 \equiv q_k$, $k = 1, \ldots, n-1$, *and* $[2(1 - s_n)/3]^3\alpha(s_n) \equiv q_n$. *Moreover, the corresponding MSE is*

$$MSE(\overline{C}_{n+1} \mid T_n, t^a_{n+1}) \sim s_1^3\alpha(s_1)/3 + \sum_{i=1}^{n-1} \alpha(s_i)d_i^3/12 + (1 - s_n)^3\alpha(s_n)/3 - \max_{0\leqslant k\leqslant n} q_k.$$

References

[1] K. Benhenni, Predicting integrals of stochastic processes: Extensions, J. Appl. Probab. 35 (1998) 843–855.
[2] K. Benhenni, S. Cambanis, Sampling designs for estimating integrals of stochastic processes, Ann. Statist. 20 (1992) 161–194.
[3] J. Istas, C. Laredo, Estimating functionals of a stochastic process, J. Appl. Probab. 29 (1997) 249–270.
[4] L.D. Pitt, R. Robeva, D.Y. Wang, An error analysis for the numerical calculation of certain random integrals: Part 1, Ann. Appl. Probab. 5 (1995) 171–197.
[5] J. Sacks, D. Ylvisaker, Designs for regression problems with correlated errors, Ann. Math. Statist. 37 (1966) 66–89.
[6] M.J. Stein, Predicting integrals of stochastic processes, Ann. Appl. Probab. 5 (1) (1995) 158–170.