BULL. AUSTRAL. MATH. SOC. 60G99, 60J30
VOL. 18 (1978), 83-93.

Estimation of the integral of a stochastic process

Noel Cressie

Consider the class of stochastic processes with stationary independent increments and finite variances; notable members are brownian motion, and the Poisson process. Now for X_t *any* member of this class of processes, we wish to find the optimum sampling points of X_t, for predicting $\int_0^A X_t dt$. This design question is shown to be directly related to finding sampling points of Y_t for estimating β in the regression equation, $Y_t = \beta t^2 + X_t$. Since processes with stationary independent increments have *linear* drift, the regression equation for Y_t is the first type of departure we might look for; namely quadratic drift, and unchanged covariance structure.

1. Introduction

Statistical inference on stochastic processes, in particular for problems of design, is an area of research which has been neglected until recently. In this paper we concentrate on processes X_t, with stationary independent increments, with finite variances, and with $X_0 = 0$; this includes both the brownian motion, and the Poisson process as special cases. Define I to be

Received 2 November 1977. The author gives due acknowledgement to F. Samaniego, who originally posed this problem.

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

$$I \equiv \int_0^A X_t dt \quad , \tag{1}$$

which is well defined as the limit in quadratic mean of the approximating Riemann sums. We are interested in the design problem of predicting this integral, based on observing the process at n distinct time points,

$$0 < t_1 < t_2 < \ldots < t_n \leq A \quad .$$

Sacks and Ylvisaker [1] studied this problem for a large class of random processes; we show here that for processes with stationary independent increments, and finite variance, the design points can be found explicitly.

The motivation for considering the prediction of $\int_0^A X_t dt$ was given by Sacks and Ylvisaker. They showed that the problem arose naturally from the following interesting question. Let Y_t be a stochastic process with $Y_0 = 0$. Now suppose Y_t can be regressed on a deterministic function $f(t)$, with X_t as error; that is

$$Y_t = \beta f(t) + X_t \quad , \quad 0 < t \leq A \quad .$$

Let $R(s, t)$ be the covariance function of X_t, and suppose

$$f(t) = \int_0^A R(s, t)\phi(s)ds \quad ,$$

ϕ continuous. Then finding the design that minimizes the variance of $\sum l_i Y_{t_i}$, the best linear unbiased estimator of β based on $\mathbf{t} = (t_1, \ldots, t_n)$, can be shown to be equivalent to finding the optimal (minimum mean squared error) design to predict $\int_0^A \phi(t)X_t dt$.

For our case $\phi(t) \equiv 1$, and we will see below that $R(s, t) = \sigma^2.\min(s, t)$. Hence the regression equation becomes

$$Y_t = \beta' t^2 + X_t' \quad ,$$

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

where $\beta' = -\beta\sigma^2/2$, and $X_t' = X_t + (A\beta\sigma^2)t$, which is still a process with stationary independent increments. In general, X_t' has a drift proportional to t (see below). Clearly then, adding the quadratic term $\beta' t^2$ to the drift is the first type of departure (from brownian motion for example) we might seek for the process Y_t . Hence finding the right design to estimate β is of primary importance. Subsequently, our interest will be in the equivalent problem of finding the optimal design to estimate $\int_0^A X_t dt$.

Define $\mathbf{t} \equiv (t_1, \ldots, t_n)$. Using minimum mean squared error as the optimality criterion, the optimal predictor becomes

$$\hat{I}_{\mathbf{t}} = E(I \mid X_{t_1}, \ldots, X_{t_n}) \tag{2}$$

$$= \int_0^A E(X_t \mid X_{t_1}, \ldots, X_{t_n})dt \ .$$

When X_t has stationary independent increments, $E(X_t \mid X_{t_1}, \ldots, X_{t_n})$ may be shown to be the polygon obtained by linearly connecting the points $(0, 0)$, (t_1, X_{t_1}), $\ldots$, (t_n, X_{t_n}) ; and for $t > t_n$, we have a straight line from (t_n, X_{t_n}) with slope m , where $m = E(X_1)$.

Also, the Lévy representation of the characteristic function of X_t tells us that

$$E(X_t X_s) = s(\sigma^2 + m^2 t) \ , \quad s < t \ ,$$

where

$$\sigma^2 = \frac{\mathrm{var}(X_t)}{t} \quad \text{and} \quad m = \frac{E(X_t)}{t} \ .$$

Therefore, $\mathrm{cov}(X_t, X_s) = \sigma^2 . \min(s, t)$.

In what is to follow, Sections 2 and 3 completely solve the design problem as posed above. Then Section 4 extends the problem to prediction

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

of $\int_u^v X_t dt$, $0 < u < v \leq A$; again the optimal design can be found explicitly.

2. To calculate the mean squared error

Define

$$\text{MSE}_t \equiv E\left[\left(I-\hat{I}_t\right)^2\right] , \tag{3}$$

where I and $\hat{I}_t$ are given in (1) and (2). Our task is to find the t which will minimize MSE_t . Firstly, we will show that without loss of generality, we may consider $m = 0$.

Define $Z_t \equiv X_t - mt$; then $E(Z_t) = 0$, for all $t \geq 0$, and

$$\begin{aligned} E(X_t \mid X_{t_1}, \ldots, X_{t_n}) &= E(X_t \mid Z_{t_1}, \ldots, Z_{t_n}) \\ &= E(Z_t \mid Z_{t_1}, \ldots, Z_{t_n}) + mt . \end{aligned}$$

Therefore,

$$X_t - E(X_t \mid X_{t_1}, \ldots, X_{t_n}) = Z_t - E(Z_t \mid Z_{t_1}, \ldots, Z_{t_n}) .$$

Hence we can do all our calculations on MSE_t as if $m = 0$; that is as if

$$E(X_t) = 0 ,$$

$$E(X_t X_s) = \sigma^2 . \min(s, t) . \tag{4}$$

We will also show that MSE_t can be partitioned into a sum of squares of orthogonal components. Consider, with $a < b < c$,

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

$$E\left\{\int_a^c [X_t - E(X_t \mid X_a, X_b, X_c)]dt\right\}^2$$

$$= E\left\{\int_a^b [X_t - E(X_t \mid X_a, X_b)]dt + \int_b^c [X_t - E(X_t \mid X_b, X_c)]dt\right\}^2$$

$$= E\left\{\int_a^b [X_t - E(X_t \mid X_a, X_b)]dt\right\}^2 + E\left\{\int_b^c [X_t - E(X_t \mid X_b, X_c)]dt\right\}^2 .$$

This is because

$$E\left\{\int_a^b [X_t - E(X_t \mid X_a, X_b)]dt \cdot \int_b^c [X_s - E(X_s \mid X_b, X_c)]ds\right\}$$

$$= \int_b^c \int_a^b E\left\{\left[X_t - \frac{X_b - X_a}{b-a}(t-a) - X_a\right]\left[X_s - \frac{X_c - X_b}{c-b}(s-b) - X_b\right]\right\}dtds$$

$$= \sigma^2 \int_b^c \int_a^b 0$$

$$= 0 .$$

Therefore, using the identical argument just given, we may write

$$(5)\quad MSE_t = E\left\{\int_0^{t_1} [X_t - E(X_t \mid X_{t_1})]dt\right\}^2 + E\left\{\int_{t_1}^{t_2} [X_t - E(X_t \mid X_{t_1}, X_{t_2})]dt\right\}^2$$

$$+ \ldots + E\left\{\int_{t_{n-1}}^{t_n} [X_t - E(X_t \mid X_{t_{n-1}}, X_{t_n})]dt\right\}^2$$

$$+ E\left\{\int_{T_n}^{A} [X_t - E(X_t \mid X_{t_n})]dt\right\}^2 .$$

We will now calculate the contribution of each of these terms to MSE_t .

Consider

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

$$E\left\{\int_a^b [X_t - E(X_t \mid X_a, X_b)]dt\right\}^2$$

$$= E\left\{\int_a^b X_t dt - \frac{(X_b+X_a)}{2}(b-a)\right\}^2$$

$$= E\left\{2\int_a^b \int_a^t X_s X_t ds dt\right\} - 2E\left\{\frac{(X_b+X_a)}{2}(b-a)\int_a^b X_t dt\right\} + E\left\{\frac{(X_b+X_a)}{2}(b-a)\right\}^2 ,$$

which, by repeated usage of (4), is equal to

$$\frac{\sigma^2(b-a)^3}{12} . \tag{6}$$

The end interval (t_n, A) is a special case. Using similar arguments to the above, we may obtain

$$E\left\{\int_{t_n}^A [X_t - E(X_t \mid X_{t_n})]dt\right\}^2 = \frac{\sigma^2(A-T_n)^3}{3} . \tag{7}$$

Therefore, (5), (6), and (7) give us

$$\mathrm{MSE}_{\mathbf{t}} = \sigma^2\left[\sum_{i=0}^{n-1} \frac{(t_{i+1}-t_i)^3}{12} + \frac{(A-t_n)^3}{3}\right] , \tag{8}$$

where $t_0 \equiv 0$.

3. Minimizing the mean squared error

We will now minimize expression (8), by putting

$$\frac{\partial}{\partial t_i}\mathrm{MSE}_{\mathbf{t}} = 0 , \quad i = 1, \ldots, n .$$

This gives us the following n equations:

$$t_i = \frac{t_{i+1}+t_{i-1}}{2} , \quad i = 1, \ldots, n-1 , \tag{9}$$

$$t_n = \frac{2A}{3} + \frac{t_{n-1}}{3} . \tag{10}$$

Solving (9) we get

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

$$t_j = \frac{j}{2} t_2 , \quad j = 1, \ldots, n ,$$

and putting this into (10) allows us to solve for t_2 . Hence the solution to (9) and (10) is

$$t_i^0 = \frac{2i}{2n+1} A , \quad i = 1, \ldots, n . \tag{11}$$

Substituting (11) into (8), we get

$$\mathrm{MSE}_{t^0} = \frac{\sigma^2 A^3}{3(2n+1)^2} ,$$

which can be shown to be greater than or equal to MSE_t for all t . Therefore, choosing $t = t^0$ gives us the optimal design.

4. Optimal design for predicting $\int_u^v X_t dt$

We now wish to find the optimal design for predicting $\int_u^v X_t dt$, where $0 < u < v \leq A$. Recall that our sampling points $t_1, t_2, \ldots, t_n$ are such that $0 < t_1 < t_2 < \ldots < t_n \leq A$. Since X_t has stationary independent increments, common sense tells us that we are going to want to sample all our points in $[u, v]$. To show that this is in fact true, suppose that we have

$$t_{j-1} \in [0, u] ,$$

and

$$t_{k+1} \in [v, A] ,$$

where $2 \leq j \leq k \leq n-1$. Define

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

$$S_t \equiv E\left\{\int_u^v [X_t - E(X_t \mid X_{t_1}, \ldots, X_{t_n})]dt\right\}^2$$

$$= E\left\{\int_u^{t_j} [X_t - E(X_t \mid X_{t_{j-1}}, X_{t_j})]dt\right\}^2$$

$$+ E\left\{\int_{t_j}^{t_{j+1}} [X_t - E(X_t \mid X_{t_j}, X_{t_{j+1}})]dt\right\}^2$$

$$+ \ldots + E\left\{\int_{t_k}^{v} [X_t - E(X_t \mid X_{t_k}, X_{t_{k+1}})]dt\right\}^2 .$$

The middle contributions to S_t are, as before,

$$\frac{\sigma^2(t_{i+1}-t_i)^3}{12}, \quad i = j, j+1, \ldots, k-1 .$$

The end contributions can be found by a method almost identical to the method used to obtain (6). They are

$$\sigma^2\left\{\frac{(t_j-u)^3}{3} - \frac{1}{4}\frac{(t_j-u)^4}{t_j-t_{j-1}}\right\}, \quad t_{j-1} \in [0, u] ,$$

and

$$\sigma^2\left\{\frac{(v-t_k)^3}{3} - \frac{1}{4}\frac{(v-t_k)^4}{t_{k+1}-t_k}\right\}, \quad t_{k+1} \in [v, A] .$$

Obviously, over all choices of t_{j-1} and t_{k+1}, the one which makes S_t the smallest is

$$t_{j-1} = u \quad \text{and} \quad t_{k+1} = v .$$

In other words, if we restrict t_{j-1} to belong to $[0, u]$, then the best choice of t_{j-1} is u, and similarly, if we restrict t_{k+1} to belong to $[v, A]$, then the best choice of t_{k+1} is v.

Common sense also tells us that a design with j as small as possible (allowing more points to be sampled in $[u, v]$), that is, with

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

$j = 2$, $k = n - 1$, will have smaller mean squared error. This can also be verified, with a little algebra.

Let us now try to minimize S_t for t satisfying

$u \le t_1 < t_2 < \ldots < t_n \le v$.

$$S_t = E\left\{\int_u^{t_1} |X_t - E(X_t \mid X_{t_1})|\,dt\right\}^2 + E\left\{\int_{t_1}^{t_2} |X_t - E(X_t \mid X_{t_1}, X_{t_2})|\,dt\right\}^2$$

$$+ \ldots + E\left\{\int_{t_n}^{v} |X_t - E(X_t \mid X_{t_n})|\,dt\right\}^2 .$$

As before, the middle contributions are

$$\frac{\sigma^2(t_{i+1}-t_i)^3}{12} , \quad i = 1, 2, \ldots, n-1 , \tag{12}$$

and from (7) the last term is

$$\frac{\sigma^2(v-t_n)^3}{3} . \tag{13}$$

The first term is derived using the method identical to the one used to get (6) and is equal to

$$\sigma^2\left[\frac{t_1^3-u^3}{3} - u^2(t_1-u) - \frac{1}{4}\frac{\left(t_1^2-u^2\right)^2}{t_1}\right] . \tag{14}$$

Add together (12), (13), (14) to give S_t , and then set

$$\frac{\partial}{\partial t_i} S_t = 0 , \quad i = 1, \ldots, n .$$

This results in the following set of n equations:

$$\begin{aligned} t_2 &= 2t_1 - u^2/t_1 , \\ t_i &= \frac{t_{i+1}+t_{i-1}}{2} , \quad i = 2, \ldots, n-1 , \\ t_n &= \frac{2v}{3} + \frac{t_{n-1}}{3} . \end{aligned} \tag{15}$$

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

Solving (15) recursively gives

$$t_1^0 = \frac{v+\sqrt{v^2+(4n^2-1)u^2}}{2n+1} , \tag{16}$$

and

$$t_j^0 = jt_1^0 - \frac{(j-1)u^2}{t_1^0} , \quad j = 1, \ldots, n . \tag{17}$$

From equation (16),

$$u^2 = \frac{\left[t_1^0\right]^2(2n+1)-2vt_1^0}{(2n-1)} ,$$

and putting this into (17) gives

$$t_j^0 = \frac{2(j-1)v}{2n-1} + \left[\frac{2n-2j+1}{2n-1}\right]t_1^0 , \quad j = 1, \ldots, n , \tag{18}$$

resulting in mean squared error

$$S_{t^0} = \sigma^2\left[\frac{\left[t_1^0\right]^3}{12} - \frac{u^2t_1^0}{2} + \frac{2}{3}u^3 - \frac{u^4}{4t_1^0}\right] + \frac{\sigma^2\left[v-t_1^0\right]^3}{3(2n-1)^2} .$$

Therefore

$$\begin{aligned} S_{t^0} &= \min\{S_t : u \le t_1, t_n \le v\} \\ &\le \min\{S_t : t_1 = u, t_n = v\} \\ &= \min\{S_t : t_1 \in [0, u], t_n \in [v, A]\} \\ &\le \min\{S_t : t_{j-1} \in [0, u], t_{k+1} \in [v, A]\} , \quad 2 \le j \le k \le n-1 , \end{aligned}$$

that is $S_{t^0} \le S_t$, for all $t \in [0, A]^n$. Hence the optimal design for predicting $\int_u^v X_t dt$, where X_t is *any* process with stationary independent increments, is given by (16) and (18).

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press

Reference

[1] Jerome Sacks and Donald Ylvisaker, "Statistical designs and integral approximation", *Time series and stochastic processes; convexity and combinatorics*, 115-136 (Proc. Twelfth Biennial Seminar, Canadian Mathematical Congress, University of British Columbia, 1969. Canadian Mathematical Congress, Société Mathématique du Canada, Montreal, Canada, 1970).

School of Mathematical Sciences,
Flinders University,
Bedford Park,
South Australia.

https://doi.org/10.1017/S0004972700007838 Published online by Cambridge University Press