BEST UNBIASED PREDICTION FOR GAUSSIAN AND LOG-GAUSSIAN PROCESSES

DENNIS D. COX*

ABSTRACT. The best linear unbiased predictor for a stochastic process is the best unbiased predictor (i.e., the linearity constraint is removed) if the process is Gaussian. This provides a stronger justification for the universal kriging predictor than is generally offered. For log-Gaussian processes, we show that the standard predictor is optimal among all unbiased predictors with respect to a weighted mean squared error prediction criterion.

1. INTRODUCTION

Consider an L_2 stochastic process $\langle Z(t) : t \in D \rangle$ (or Z for short) with mean function of the form

$$E[Z(t)] = \sum_{j=1}^{p} \beta_j B_j(t),$$

and covariance function

$$\mathrm{Cov}[Z(s), Z(t)] = \sigma^2 K(s,t).$$

We assume the basis functions $B_j(t)$, $1 \leq j \leq p$, are known but the coefficients β_j, $1 \leq j \leq p$ are unknown. The covariance $\sigma^2 K(s,t)$ is assumed known except for possibly the scale parameter σ^2. This is essentially the setup of universal kriging (or ordinary kriging if $p = 1$ and $B_1(t) \equiv 1$). See Cressie (1991). Here we consider prediction of a value $Z(s)$ given values $(Z(t_1), \ldots, Z(t_n))$, which is the usual geostatistical prediction problem. The universal kriging predictor is what is typically used in such prediction problems. Besides the geostatistical applications, it has been used elsewhere for prediction (or approximation, interpolation, or smoothing) of general functions (Sacks, et. al., 1989)

*This research was supported by National Science Foundation grant DMS-0204723.

Key Words and Phrases: prediction, kriging, Gaussian processes, log-Gaussian processes, unbiasedness, unbiased prediction.

based on a "random function" model (or function space prior, if one is willing to admit to the Bayesian nature of such models).

The universal kriging predictor is typically justified on the grounds that it is "best" among linear predictors which are "unbiased," i.e. it is the "best linear unbiased predictor" or BLUP, where, "best" is in the sense of minimizing the expected value of the squared prediction error. Definitions are given below. Here, we show that if the process is Gaussian, then it is in fact best among all unbiased predictors, i.e. the "best unbiased predictor" or BUP. While the Gaussian assumption is rather restrictive, it is in fact the assumption of choice when actual distributions are needed rather than just first and second moments (e.g. for computation of prediction intervals, likelihoods, etc.), and it is comforting to know that the widely used predictor has the additional optimality property (just as one is perhaps comforted to know that ordinary least squares estimator in regression is UMVUE under a Gaussian errors model as well as BLUE for the Gauss-Markov setup; see Lehmann, 1983). Furthermore, the result here provides an alternate proof to the fact that the BUP for the Gaussian process is BLUP in general (see Corollary 3.3 below).

When the process is clearly not Gaussian, it is often assumed that some transformation makes it Gaussian, and the most commonly used Gaussianizing transformation is probably the logarithm. A process Z whose logarithm is Gaussian is called log-Gaussian. Now for such log-Gaussian processes, there is a well developed prediction theory, the so-called log-normal kriging. See section 3.2.2 of Cressie (1991). Assuming σ^2 is known, one constructs an unbiased predictor of a log-Gaussian process by exponentiating the BLUP (now known to be BUP) of the underlying Gaussian process and multiplying by a suitable constant to correct for bias. As in the case of a Gaussian process, we show here that this predictor is optimal among all unbiased predictors (not just those obtained by exponentiating a linear predictor based on the underlying Gaussian process), but the optimality is in terms of a weighted mean squared prediction error.

The results concerning the optimality of the BLUP for Gaussian processes are certainly not surprising. As mentioned in Handcock and Stein (1993), the BLUP is the posterior mean predictor under a uniform prior on β (assuming K is known). The result for the log-Gaussian process is perhaps of greater interest (although we have adjusted the loss function, but not in an unreasonable way). We find the method of proof to be most intriguing. As in the theory of classical unbiased estimation, an important role is played by the completeness and sufficiency of the statistic $\hat{\beta}$, the generalized least squares estimator of the regression coefficient vector. It would be of interest to know if other prediction problems are amenable to these techniques.

The next section sets up the mathematical framework for the main results, which is in fact more general and simpler than a stochastic process. In the third section are given the theorem and proof for the Gaussian case, and in the fourth the result for the log-Gaussian case.

2. Mathematical Framework

While we have expressed the problem above in terms of stochastic processes, it of course can be reduced to one involving $(n+1)$ dimensional normal random vectors. Let X be a random n dimensional vector and Y a random variable so that (X, Y) has a multivariate normal distribution on $\mathbb{R}^{n+1}$. For the Gaussian process case, $X = (Z(t_1), \ldots, Z(t_n))$ and $Y = Z(s)$. Vectors are represented as column matrices or ordered n-tuples. Assume that the mean is given by

$$E\left\{\begin{bmatrix} X \\ Y \end{bmatrix}\right\} = \begin{bmatrix} \xi \\ \eta \end{bmatrix}\beta \tag{2.1}$$

where ξ is a known $n \times p$ matrix, η is a known $1 \times p$ matrix, and β is an unknown p vector. We assume ξ is of full column rank p. Denoting the (i, j) entry of a matrix B by $B[i, j]$, of course, $\xi[i, j] = B_j(t_i)$ and $\eta[1, j] = B_j(s)$. The covariance of (X, Y) is written in partitioned form as

$$Cov\left\{\begin{bmatrix} X \\ Y \end{bmatrix}\right\} = \sigma^2 \begin{bmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{bmatrix}, \tag{2.2}$$

where K_{XX} is $n \times n$, $K_{XY} = K_{YX}^T$ is $n \times 1$, and K_{YY} is 1×1. Of course $K_{XX}[i, j] = K(t_i, t_j)$, $K_{XY}[i, 1] = K(t_i, s)$, and $K_{YY}[1, 1] = K(s, s)$, where K is the covariance function given above. We assume K_{XX} is nonsingular.

In the Gaussian case, we consider predictors $p(X)$ of Y which are a function of X, i.e. a predictor of Y given X. Such a predictor is called unbiased if

$$E[p(X)] = E[Y] = \eta\beta, \tag{2.3}$$

for all $\beta \in I\!R^p$. We will consider only unbiased predictors as defined here. This restriction is not as easy to justify as unbiasedness in point estimation as considered in Lehmann (1983), but it has been widely accepted. We note that in the stochastic process setting, the sampling design $t_1, t_2, \ldots, t_n$ is treated as fixed, so if the design is in fact random, then this notion of unbiasedness is the analogue of conditional unbiasedness as discussed in Shaffer (2000) for estimation. It may be of interest to consider unconditional unbiasedness in this setting to see if it leads to the discovery of other predictors.

For any predictor $p(X)$ the mean squared prediction error is given by

$$MSPE[p] = E\left[(p(X) - Y)^2\right]. \tag{2.4}$$

One unbiased predictor is

$$p_0(X) = \eta\hat{\beta} + A\left(X - \xi\hat{\beta}\right), \tag{2.5}$$

where

$$\hat{\beta} = \left(\xi^T K_{XX}^{-1}\xi\right)^{-1}\xi^T K_{XX}^{-1} X,$$

is the generalized least squares estimator of β and

$$A = K_{YX}K_{XX}^{-1}. \tag{2.6}$$

One can obtain $p_0(X)$ informally by starting with the formula for $E[Y|X]$ and plugging in $\hat{\beta}$ for the unknown β.

For the log-Gaussian case, $X = (\log Z(t_1), \ldots, \log Z(t_n))$ and $Y = \log Z(s)$, and we wish to predict $W = Z(s) = \exp Y$. For this case, we assume σ^2 is known. We consider predictors $q(X)$ of W which are a function of X, and such a predictor is called unbiased if

$$\begin{aligned} E[q(X)] &= E[W] = \exp\{E[Y] + (1/2)\mathrm{Var}[Y]\} \\ &= \exp[\eta\beta + (1/2)\sigma^2 K_{YY}], \end{aligned} \tag{2.7}$$

for all $\beta \in I\!R^p$. For any predictor $q(X)$ we consider the weighted mean squared prediction error given by

$$WMSPE[p] = E\left[\exp[-AX]\,(q(X) - Y)^2\right], \tag{2.8}$$

where A is given in (2.6). This is clearly equivalent to

$$WMSPE_2[p] = E\left[(q(X) - Y)^2 / E[W|X]\right],$$

as

$$\begin{aligned} E[W|X] &= \exp\{E[Y|X] + (1/2)\mathrm{Var}[Y|X]\} \\ &= \exp\left[\eta\beta + A(X - \xi\beta) + (1/2)\left(K_{YY} - AK_{XY}\right)\right]. \end{aligned}$$

One such unbiased predictor is

$$q_0(X) = C\exp\left[p_0(X)\right], \tag{2.9}$$

where $p_0(X)$ is the BUP of Y given in (2.5) and C corrects for bias:

$$\begin{aligned} C &= \exp\left[(1/2)\left\{\mathrm{Var}[Y] - \mathrm{Var}[p_0(X)]\right\}\right] \\ &= \exp\left[(1/2)\sigma^2\left\{K_{YY} - \left(K_{YY} - AK_{XY} - \eta(\xi^T K_{XX}^{-1}\xi)^{-1}\eta^T\right.\right.\right. \\ &\qquad \left.\left.\left. + A\xi(\xi^T K_{XX}^{-1}\xi)^{-1}\xi^T A^T\right)\right\}\right] \\ &= \exp\left[(1/2)\sigma^2\left\{AK_{XY} + \eta(\xi^T K_{XX}^{-1}\xi)^{-1}\eta^T - A\xi(\xi^T K_{XX}^{-1}\xi)^{-1}\xi^T A^T\right\}\right] \end{aligned} \tag{2.10}$$

It follows from Theorem 4.1 below that this is the "best" unbiased predictor of W given X in the sense of minimizing WMSPE.

3. Best Unbiased Prediction for Gaussian Processes

Here we present the theorem which justifies our claim that the BLUP is in fact BUP for Gaussian processes.

Theorem 3.1. *Among all unbiased predictors of Y given X, $p_0(X)$ in (2.5) minimizes MSPE.*

Proof. First of all note that $\hat{\beta}$ is an unbiased estimator of β, from which it follows that $p_0(X)$ is an unbiased predictor of Y given X. Let $p(X)$ be any other unbiased predictor of Y given X. Then

$$\begin{aligned} MSPE[p] &= E\left[(Y - p_0(X))^2\right] + E\left[(p_0(X) - p(X))^2\right] \\ &\quad + 2E\left[(Y - p_0(X))\left(p_0(X) - p(X)\right)\right]. \end{aligned}$$

The theorem follows once we show

$$E\left[(Y - p_0(X))\left(p_0(X) - p(X)\right)\right] = 0. \tag{3.1}$$

By elementary properties of conditional expectation,

$$\begin{aligned} & E\left[(Y - p_0(X))\,(p_0(X) - p(X))\right] && (3.2)\\ &= E\left\{(p_0(X) - p(X))\,E\left[(Y - p_0(X))\,|X\right]\right\} \\ &= E\left[(p_0(X) - p(X))\,(\eta\beta + A(X - \xi\beta) - p_0(X))\right] && (3.3)\\ &= (\eta - A\xi)E\left[(p_0(X) - p(X))\left(\beta - \hat\beta\right)\right]. && (3.4)\end{aligned}$$

In the above, (3.3) follows from the well known formula for conditional expectations for one component of a multivariate normal distribution given the rest, and (3.4) follows from the definition of $p_0(X)$ and a little algebra. Continuing to calculate with the last expression,

$$E\left[(p_0(X) - p(X))\left(\beta - \hat\beta\right)\right] = E\left\{\left(\beta - \hat\beta\right)E\left[p_0(X) - p(X)|\hat\beta\right]\right\}. \quad (3.5)$$

Now if σ^2 is known, then $\hat\beta$ is a complete sufficient statistic for β by Lemma 3.2 below. Hence, by sufficiency, $h(\hat\beta) = E\left[p_0(X) - p(X)|\hat\beta\right]$ does not depend on β, i.e. is a statistic. Furthermore, for any value of β, unbiasedness of $p(X)$ and $p_0(X)$ implies $E[h(\hat\beta)] = 0$, so by completeness, $h(\hat\beta) = 0$ almost surely. Plugging this into (3.5) and then into the previous calculations establishes (3.1).

Since $p_0(X)$ does not depend on σ^2, it follows that it is the unbiased predictor of Y given X which minimizes MSPE both when σ^2 is known and when it is not known. This completes the proof.

Lemma 3.2. *With σ^2 known, $\hat\beta$ is a complete sufficient statistic for β.*

Proof. The result is presumably well known, but we could not find a quick reference to it. One may write the normal likelihood in exponential family form as

$$\begin{aligned} f(x|\beta) &= g(\beta)\exp\left[(1/\sigma^2)\beta^T\xi^T K_{XX}^{-1}x\right]h(x) \\ &= g(\beta)\exp\left[(1/\sigma^2)\{(\xi^T K_{XX}^{-1}\xi)\beta\}^T\hat\beta\right]h(x),\end{aligned}$$

where $g(\beta)$ does not depend on x and $h(x)$ does not depend on β. We need only verify the full rank condition of Lehmann (1983), page 28 (see Theorem 5.6, page 46). To this end, note that $\hat\beta$ satisfies no linear constraints (it ranges over all of $\mathbb{R}^p$ as x ranges over $I\!R^n$) and $(\xi^T K_{XX}^{-1}\xi)\beta$ ranges over all of $I\!R^p$ as β ranges over $I\!R^p$. This completes the proof.

A simple corollary is the usual result that $p_0(X)$ is the BLUP for a general process.

Corollary 3.3. *Suppose X is a random n dimensional vector and Y is a random variable both having finite second moments with mean and covariance given by (2.1) and (2.2), respectively. Then $p_0(X)$ minimizes MSPE among all linear unbiased predictors of Y given X.*

Proof. Clearly $p_0(X)$ is a linear unbiased predictor of Y given X. The MSPE of a linear unbiased predictor depends only on its variance and covariance with Y, so is the same whether the process is Gaussian or not. The proof follows from these simple facts and the Theorem.

4. A Result for Log Gaussian Processes

Here we state and prove the theorem which justifies the optimality claim for the unbiased predictor of a log-Gaussian process. Recall that σ^2 is assumed known for this case.

Theorem 4.1. *Among all unbiased predictors of $W = \exp[Y]$ given X, $q_0(X)$ in (2.9) minimizes WMSPE.*

Proof. One can check the $q_0(X)$ is an unbiased predictor. Let $q(X)$ be any other unbiased predictor of W given X. First note that

$$\begin{aligned} E[W|X] &= \exp[E[Y|X] + (1/2)\text{Var}[Y|X]] \\ &= \exp[\eta\beta + A(X-\xi\beta) + (1/2)\text{Var}[Y|X]], \end{aligned}$$

where $\text{Var}[Y|X]$ is a constant not depending on X. Hence,

$$\begin{aligned} E[q_0(X) - W|X] &= C\exp\left[\eta\hat{\beta} + A(X-\xi\hat{\beta})\right] - \\ &\qquad \exp[\eta\beta + A(X-\xi\beta) + (1/2)\text{Var}[Y|X]] \\ &= e^{AX}\exp\left[(\eta - A\xi)\hat{\beta}\right]\left(C - C'\exp\left[(\eta-A\xi)(\beta-\hat{\beta})\right]\right), \end{aligned}$$

with $C' = \exp[(1/2)\mathrm{Var}[Y|X]]$. Following the argument of Theorem 3.1, we jump in at the analog of (3.2):

$$
\begin{aligned}
&E\left[e^{-AX}(W - q_0(X))(q_0(X) - q(X))\right] \\
&= -E\left[\exp\left[(\eta - A\xi)\hat{\beta}\right]\left(C - C'\exp\left[(\eta - A\xi)(\beta - \hat{\beta})\right]\right)(q_0(X) - q(X))\right] \\
&= -E\left[\exp\left[(\eta - A\xi)\hat{\beta}\right]\left(C - C'\exp\left[(\eta - A\xi)(\beta - \hat{\beta})\right]\right)E\left[q_0(X) - q(X)|\hat{\beta}\right]\right].
\end{aligned}
$$

As before, it follows from the completeness and sufficiency of $\hat{\beta}$ and the fact that $q_0(X)$ and $q(X)$ have the same expectations as a function of β that $E[q_0(X) - q(X)|\hat{\beta}] = 0$, a.s. and the result follows.

Remark: While we adjusted the loss function to make the proof of the previous theorem go through, it is clear that one could retain the unweighted squared error loss and modify the definition of unbiasedness and obtain a result.

Dennis D. Cox
Department of Statistics
P.O. Box 1892
Rice University
Houston, TX, 77251

dcox@stat.rice.edu