Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

PREDICTION FUNCTIONS AND GEOSTATISTICS

A. F. Desmond
University of Guelph

ABSTRACT

We consider analogues of estimating functions for situations in which the prediction of observables is of primary interest. We show that the mode of the predictive density has an optimality property for prediction analogous to a similar optimality property for the mode of the posterior density in the case of parametric estimation. Applications of predictive estimating functions in spatial statistics with particular reference to the geostatistical method known as kriging are developed.

Key Words: Predictive estimating functions; kriging; predictive density; spatial statistics.

1 Introduction

Applications of estimating functions have mainly focused on estimation and inference for parameters either in fully parametric or semi-parametric models. While the focus on parameters as indices of probability distributions is at the core of modern statistics ever since Fisher (1922) (See Stigler (1976)), several authors have argued persuasively that prediction of potential observables may sometimes be important. For example, Pearson (1920) refers to this as the Fundamental Problem of Statistics. It is also central to the treatment of de Finetti (1974, 1975) although he prefers the term prevision. Geisser (1993) gives an extensive discussion of predictive inference focusing, however, mainly on the Bayesian approach. In this article we discuss predictive analogues of estimating functions, motivated by similar ideas for parametric estimation. Such analogues might be termed prediction functions, although this term has been used previously in at least two different senses. Mathiasen (1979) uses it to denote a function of the future observable and the current data which ranks values of the future observation in terms of relative plausibility in the light of the data. Such a function has also

been termed a predictive likelihood (cf. Bjornstad (1990) for a comprehensive review). The term prediction function or predictor has also been used for a function of the current data used to predict a future observation or a function of several future observations, e.g. in time series. Here, I shall use the term prediction function or predictive estimating function (to emphasize the analogy with estimating functions) to denote a function $g(z, \underline{x})$ of data $\underline{x}$ and an observable z to be predicted. The equation

$$g(z, \underline{x}) = 0$$

will be referred to as a predictive estimating equation.

In this article I consider predictive inference from the point of view of prediction functions. In section three I consider this from the Bayesian point of view and show that one can obtain optimality properties of modal predictive density estimators based on two optimality criteria analogous to those of Godambe (1960), Ferreira (1982), and Ghosh (1990). These optimality criteria are introduced in section two. In sections four and five I consider applications to spatial prediction and show that the kriging equations of Matheron (1962) can be obtained as special cases. Section six deals with prediction functions related to predictive likelihood, while section seven concludes with a brief discussion.

2 Optimality Criteria for Prediction

Classical approaches to prediction in time series (e.g. Box and Jenkins (1970)) or spatial statistics (Ripley (1981)) invariably focus on mean-squared error as a criterion e.g. in the 1-step ahead prediction problem one considers $E[Y_f - h(Y_c)]^2$ either unconditionally or conditionally on Y_c, where Y_f denotes a future observation, and $h(Y_c)$ is a function of current data Y_c.

Suppose we are interested in a future value z from a parametric family $f_z(z; \underline{\theta})$ depending on an unknown parameter $\underline{\theta}$. Denote the current data by $\underline{x} = (x_1, \ldots, x_n)$ a random sample from $f_x(x; \underline{\theta})$. An unbiased prediction function is a function g of the current data $\underline{x}$ and a future observable z such that:

$$E_{(z,\underline{x})}\{g(z, \underline{x})\} = 0 \tag{2.1}$$

the expectation being over the joint distribution of z and $\underline{x}$. We could also consider conditional unbiasedness:

$$E_{z/\underline{x}}\{g(z, \underline{x})\} = 0. \tag{2.2}$$

There is some debate (e.g. Butler (1990)) concerning the appropriateness of unconditional versus conditional assessments of predictive inferences but in my view both measures are important in practical applications. Roughly

speaking, though, it is probably the case that unconditional assessments are more relevant in pre-data considerations, while conditional assessment may be more relevant at the analysis stage.

An important difference here with more conventional treatments of prediction is that the point predictor $h(\underline{x})$, say, obtained as a solution in z of the equation

$$g(z, \underline{x}) = 0 \tag{2.3}$$

may or may not be unbiased. (A sufficient condition for the conventional unbiasedness requirement, $E(z) = E(h(\underline{x}))$, is linearity of g in z and $\underline{x}$).

Another departure from classical prediction problems we propose to adopt here is to define optimality of our prediction function in terms of one of the following optimality criteria:

$$EFF(g) = \frac{E_{(z,\underline{x})}(g^2)}{\{E_{(z,\underline{x})}(\frac{\partial g}{\partial z})\}^2} \tag{2.4}$$

or

$$EFF(g) = \frac{E_{(z|\underline{x})}(g^2)}{\{E_{z|\underline{x}}(\frac{\partial g}{\partial z})\}^2} \tag{2.5}$$

where EFF denotes efficiency.
For simplicity we restrict attention to scalar z although multivariate analogues of (2.4) and (2.5) are easily obtained. Also, (2.4) and (2.5) are predictive analogues of criteria previously proposed by Ferreira (1982) and Ghosh (1990) respectively. Those authors, however, are concerned with Bayesian estimation of unknown and unobservable parameters, whereas z here represents a random variable which is potentially observable exactly (i.e. without measurement error). Naik-Nimbalkar and Rajarshi (1995) also develop this framework extensively in the context of state-space models but again the state-variables are parameters which are unobservable (although allowing measurement with error).

In this paper, I shall restrict attention to the unconditional criteria, (2.1) and (2.4), although a similar development is possible in terms of the conditional criteria.

3 Optimal Prediction Functions

Suppose z is a scalar future observable and $\underline{x} = (x_1, \ldots, x_n)$ is the current data. The ideal object for predictive purposes is a conditional probability density of the future observable z given the current data $\underline{x}$, $p(z|\underline{x})$ say. (We assume that all random variables are continuous). Lindley (1990) and Geisser (1993) point out that full specification of the marginal density of $\underline{x}$, $p(\underline{x})$ say, and hence, a fortiori, $p(y|\underline{x})$, is difficult in general, but that

the introduction of a parametric model $q(\underline{x}|\theta)$ and a prior distribution $\pi(\theta)$ enables us to write

$$p(\underline{x}) = \int q(\underline{x}|\theta)\pi(\theta)d\theta$$

and, hence, the predictive density can be calculated. Also, the de Finetti representation theorem (de Finetti (1937)) shows that if the x_i are exchangeable, such a "lurking" parametric structure is implied. If one accepts the Bayesian argument, the predictive density can be obtained as

$$p(y|\underline{x}) = \int p(y|\underline{x}), \theta\pi(\theta|\underline{x})d\theta \tag{3.1}$$

where $\pi(\theta|\underline{x})$ is the posterior density of θ given $\underline{x}$, calculated from Bayes theorem. Fisher (1956) and Kalbfleisch (1971) obtained predictive distributions in situations where a fiducial distribution for the parameters is available. The calculation is similar to (3.1) with the posterior density replaced by the fiducial density, although the logic is quite different. Aitchison and Dunsmore (1975) make extensive use of the Bayesian predictive density.

We argue here, informally, that if a predictive density is available, then the optimal prediction function is given by:

$$g^* = \frac{\partial lnp(z|\underline{x})}{\partial z}. \tag{3.2}$$

We assume regularity conditions similar to those of Ferreira (1982), but involving conditions on existence of certain derivatives with respect to z rather than θ, and that certain interchanges of differentiation and integration operators are permissible. Denote by G the class of all prediction functions $g : Z \times X \to \Re$, such that $E_{\underline{x}}E_{z|\underline{x}}(g) = 0$ and $E_{\underline{x}}E_{z|\underline{x}}(g^2) < \infty$, where Z and X are the obvious sample spaces. Then G is a vector space with respect to the real numbers in the sense that if c_1, c_2 are real numbers and $g_1, g_2 \in G$, then $c_1g_1 + c_2g_2 \in G$. Also, if we define the inner product $< g_1, g_2 >= E_{\underline{x}}E_{z|\underline{x}}(g_1g_2)$, G is an inner product space which is complete in the metric defined by the norm $||g||^2 =< g, g >$; hence G is also a Hilbert space.

Consider an arbitrary element g of G. Denote by $\dot{g}$ the derivative of g with respect to z. Differentiation under the integral sign in the condition:

$$E_{\underline{x}}E_{z|\underline{x}}(g) = 0,$$

and arguing in a similar fashion to Ferreira (1982) yields:

$$E(\dot{g}) = - < g, g^* >$$

where

$$g^* = \frac{\partial lnp(z|\underline{x})}{\partial z}$$

is the logarithmic derivative of the predictive density with respect to the future observable z. The Cauchy-Schwarz inequality now implies

$$[E(\dot{g})]^2 = | < g, g^* > |^2 \leq ||g||^2 ||g^*||^2.$$

Optimality of g^* then follows from the condition for equality in the Cauchy-Schwarz inequality.

In practice, the predictive density $p(z|\underline{x})$ will be unavailable and it is necessary to restrict attention to subclasses of unbiased prediction functions. For example, if G_1 is the class of prediction functions linear in z and $\underline{x}$, standard arguments suggest that the optimal prediction function in G_1 is the projection of g^* above into G_1. Minimization of (2.4) above is then equivalent to finding $g_1^* \in G_1$ such that

$$||g_1^* - g^*|| \leq ||g_1 - g^*||, \quad \text{for all} \quad g_1 \in G_1.$$

We give an example of this in spatial statistics in the next section and show that an optimal g_1^* can be found depending only on second-moment assumptions about the data and the "future" observable.

I make two remarks at this point. Firstly, by appropriate redefinition of inner products and corresponding norms, the above development can be carried out mutatis mutandis in terms of the conditional quantities (2.2) and (2.5). Secondly, optimality of the mode of the predictive density can also be derived based on maximizing an expected utility function. Aitchison and Dunsmore (1975, p. 46) show that for an all-or-nothing utility structure the optimum point predictor is the mode of the predictive density.

4 Prediction Function Approaches to Kriging

Godambe (1985) investigated finite sample parametric estimation for stochastic processes using estimating functions. The stochastic processes he considered were in discrete time and optimal quasi-score functions based on elementary martingale estimating functions were constructed. Thavaneswaran and Thompson (1986) generalize this to continuous time processes. In the case of spatial statistics such a martingale formulation would appear to be inapplicable. Nevertheless it is of interest to consider what estimating functions may offer. Also, unlike the aforementioned authors who deal with fixed unknown parameters in stochastic models, the type of applications I consider here involve prediction of unobserved (but potentially observable) random variables.

We are particularly concerned here with an approach to spatial prediction commonly referred to as kriging, which has found extensive application in such areas as hydrology, soil science, and the mining industry (e.g. Journel and Huijbregts (1978)). Kriging is, in essence, an analogue for spatial

processes of the optimal linear prediction theories of Kolmogorov (1941) and Wiener (1949) for time-series and was developed mainly by Matheron and his school in the mining industry. Cressie (1990) gives an interesting historical account of the origins of kriging. As the area is relatively unfamiliar to statisticians we outline briefly some of the main ideas. We assume an underlying two or three-dimensional spatial stochastic process $z(\underline{x})$, $\underline{x} \in \Re^2$ or $\Re^3$ representing, for example, soil PH in $\Re^2$ or ore-grade in $\Re^3$. It is desired to estimate or predict $z(\underline{x}_0)$ at some unobserved location $\underline{x}_0$, based on observations of $z(\underline{x})$, $z(\underline{x}_1), \ldots, z(\underline{x}_n)$ at a set of n spatial locations $\underline{x}_1, \ldots, \underline{x}_n$. The optimal predictor in terms of minimizing the mean-squared prediction error

$$E[\hat{z}(\underline{x}_0; z(\underline{x}_1), ..., z(\underline{x}_n)) - z(\underline{x}_0)]^2 \tag{4.1}$$

is, as in the Wiener-Kolmogorov theory, given by, the conditional expectation

$$E(z(\underline{x}_0)|z(\underline{x}_1), \ldots, z(\underline{x}_n)) \tag{4.2}$$

but this entails knowledge of an $(n+1)$-dimensional distribution which may not be available. Kriging focuses on linear predictors of the form

$$\hat{z}(\underline{x}_0; z(\underline{x}_1), \ldots, z(\underline{x}_n)) = \sum_{i=1}^{n} \lambda_i z(\underline{x}_i) + \lambda_0 \tag{4.3}$$

which satisfy an unbiasedness condition

$$E[\hat{z}(\underline{x}_0; z(\underline{x}_1), \ldots, z(\underline{x}_n)) - z(\underline{x}_0)] \equiv 0, \tag{4.4}$$

and seeks a predictor of the form (4.3) which minimizes (4.1) subject to (4.4).

The solution to the kriging problem, i.e. determination of the optimal weights $\lambda_1, \lambda_2 \ldots \lambda_n$, depends on assumptions about the structure of $z(\underline{x})$. Usually, it is assumed that

$$z(\underline{x}) = \theta(\underline{x}) + \epsilon(\underline{x})$$

where $\epsilon(\underline{x})$ has zero mean and is either a stationary process or an intrinsic random function (Matheron (1962)). An intrinsic random function is one for which generalized increments of some order are second-order stationary (e.g. Cressie (1991, p. 300)). There are three common assumptions about $\theta(\underline{x})$: (i)$\theta(\underline{x})$ is a known constant θ_0, (ii) $\theta(\underline{x})$ is an unknown constant θ_0 and (iii) $\theta(\underline{x})$ is a "trend" function of the form $\sum_{j=1}^{p} f_j(\underline{x})\beta_j$ where $f_j(\underline{x})$ are known functions (e.g. low-order polynomials) and β_j are unknown parameters. Under assumptions of known covariance structure $\Sigma(\underline{x}_1 - \underline{x}_2) = Cov(\epsilon(\underline{x}_1), \epsilon(\underline{x}_2))$, in the case where $\epsilon(\underline{x})$ is stationary, or known semivariogram $\gamma(\underline{x}_1 - \underline{x}_2) = \frac{1}{2}E[(\epsilon(\underline{x}_1) - \epsilon(\underline{x}_2))^2]$ in the case where $\epsilon(\underline{x})$ is an intrinsic random

function of order zero, optimal predictions in terms of (4.1) can be obtained in cases (i), (ii) and (iii) and lead, respectively, to what are referred to as simple, ordinary or universal kriging equations for the coefficients λ_i in (4.3). Details are given in Ripley (1981) or Cressie (1991). Optimality here rests on the strong assumption of known covariance function or semi-variogram. In practice, this needs to be estimated and an enormous body of work in geostatistics has concentrated on its estimation. Often various simple parametric forms for $\gamma_\theta()$ are assumed, θ estimated in various ad hoc ways, and $\gamma_{\hat{\theta}}(\cdot)$ is inserted into the kriging equations.

Consider now a reformulation of the kriging problem from a predictive estimating function point of view. We seek a prediction function $g(z(\underline{x}_0), z(\underline{x}_1), \ldots, z(\underline{x}_n)$ which is: (i) unbiased,

$$E\{g(z(\underline{x}_0), z(\underline{x}_1), \ldots, z(\underline{x}_n))\} = 0 \tag{4.5}$$

and (ii) minimizes

$$\frac{Eg^2}{[E(\frac{\partial g}{\partial z})]^2} \tag{4.6}$$

where $\frac{\partial g}{\partial z}$ is the derivative of g evaluated at $z(\underline{x}_0) = z$. Clearly (4.4) is a special case of (4.5) but (4.5) is more general in that the predictor $\hat{z}(\underline{x}_0)$ which is a solution to $g = 0$ need not be unbiased in the sense of (4.4).

From section 3, the optimal predictor, minimizing (ii) in the class of unbiased predictive functions is

$$g^* = \frac{\partial lnp(z(\underline{x}_0) = z | z(\underline{x}_1), \ldots, z(\underline{x}_n))}{\partial z} \tag{4.7}$$

provided the necessary conditional distribution is available. This is unlike the Wiener-Kolmogorov-Matheron theory which leads to the conditional expectation predictor. It has the same difficulties, which that theory encounters, in that knowledge of the conditional distribution is rarely available. However, if $z(\underline{x})$ is a stationary Gaussian process, the modal predictor according to (4.7) coincides with the conditional expectation predictor.

To obtain a prediction function not predicated on strong distributional assumptions we restrict the class of competing prediction functions. For example, one possibility is the class:

$$\begin{aligned} G_1 &= \{g : g = g(z(\underline{x}_0), z(\underline{x}_1), \ldots, z(\underline{x}_n)) \\ &= g_1(z(\underline{x}_0) - h(z(\underline{x}_1), \ldots, z(\underline{x}_n))) \text{ and } E(g_1) = 0\} \end{aligned} \tag{4.8}$$

where g_1 and h are possibly non-linear functions. Clearly, choice of the identity function for g_1 is sensible in many applications. For example, the

disjunctive kriging of Matheron (1976) would correspond to g_1 the identity function and

$$h(z(\underline{x}_1), \ldots, z(\underline{x}_n)) = \sum_{i=1}^{n} h_i(z(\underline{x}_i))$$

where the functions $\{h_i : i = 1, ..., n\}$ are measurable square-integrable functions. Matheron (1976) shows that minimum mean-squared prediction error predictors can be obtained. The resulting disjunctive kriging equations require knowledge only of the bivariate distributions of $(z(\underline{x}_i), z(\underline{x}_j))$, provided the process $z(\cdot)$ follows a so-called isofactorial model. This is in contrast to (4.7) which typically involves knowledge of an $(n+1)$-dimensional distribution.

A further special case of (4.8) is the class

$$\begin{aligned} G_2 &= \{g : g = g(z(\underline{x}_0), z(\underline{x}_1), ..., z(\underline{x}_n)) \\ &= z(\underline{x}_0) - \lambda_0 - \sum_{i=1}^{n} \lambda_i z(\underline{x}_i), \quad \text{with} \quad E(g) = 0\} \end{aligned} \tag{4.9}$$

corresponding to prediction functions linear in the observations $z(\underline{x}_i)$, $1 \leq i \leq n$ and the unobserved $z(\underline{x}_0)$.

We now show that the optimal prediction function in the class G_2 leads to the simple kriging equations of Matheron (1962). The argument is a modification of theorem 2.1 of Thavaneswaran and Thompson (1988). We give the result for second-order stationary processes with known covariance function although it can be modified for the intrinsic random function situation with known semi-variogram.

Theorem 4.1. Suppose $E(z(\underline{x})) = \theta(\underline{x})$ is a known function and denote by Σ_{ZZ} the $n \times n$ matrix with (i,j)th element $Cov(z(\underline{x}_i), z(\underline{x}_j))$. Let $\theta_0 = E(z(\underline{x}_0))$ and, $\underline{\theta} = (\theta(\underline{x}_1), \ldots, \theta(\underline{x}_n))^T$ and $\underline{d}$ be the n-vector with ith element $Cov(z(\underline{x}_0), z(\underline{x}_i))$. Let G_2 be as in (4.9) above rewritten in the form

$$G_2 = \{g : g = (z(\underline{x}_0) - \theta_0) - \lambda^T(\underline{z}_n - \underline{\theta})\}$$

where $\lambda^T = (\lambda_1, ..., \lambda_n)$ and $\underline{z}_n = (z(\underline{x}_1), ..., z(\underline{x}_n))^T$. Then the optimal prediction function minimizing (4.6) is given by

$$g^* = (z(\underline{x}_0) - \theta_0) - \underline{d}^T \Sigma_{ZZ}^{-1}(\underline{z}_n - \underline{\theta}). \tag{4.10}$$

Proof. The proof is an elementary modification of that of Thavaneswaran and Thompson (1988). They point out that a sufficient condition for g^* to be optimal is that $E(gg^*) = KE(\dot{g})$ where $\dot{g}$ is the derivative with respect to $z(\underline{x}_0)$ and K is an arbitrary constant. Since for the class G_2, $E(\dot{g})$ is unity,

it suffices to show that $E(gg^*) = E(g^{*2})$ for all $g \in G_2$. For g^* given by (4.10), elementary manipulation yields

$$E(g^*g) = Var(z(\underline{x}_0)) - \underline{d}^T\Sigma_{ZZ}^{-1}\underline{d}, \tag{4.11}$$

for all $g \in G_2$. Since the right hand side does not involve $\underline{\lambda}$, and hence not on the choice of g, the result is proven.

In this case the optimal predictor $\hat{z}(\underline{x}_0)$ obtained by solving $g^* = 0$ for $z(\underline{x}_0)$ is given by

$$\hat{z}(\underline{x}_0) = \theta_0 + \underline{d}^T\Sigma_{ZZ}^{-1}(\underline{z}_n - \underline{\theta}). \tag{4.12}$$

Matheron's derivation involves minimizing the mean-squared error with respect to variation in $\lambda_0, \lambda_1, ..., \lambda_n$, subject to $E(\hat{z}(\underline{x}_0)) = E(z(\underline{x}_0))$. This leads via a Lagrange multiplier construction to the $n+1$ linear equations for $\lambda_0, \lambda_1, \ldots, \lambda_n$ given by:

$$\begin{aligned} \lambda_0 + \lambda^T\underline{\theta} &= \theta_0 \\ \lambda_0\theta(x_i) &+ \lambda^T\Sigma_{Z_iZ} = \underline{d}, i =, ...n \end{aligned}$$

where Σ_{Z_iZ} is the n-vector whose jth element is $Cov(z(\underline{x}_i), z(\underline{x}_j))$. The optimal weights are identical to those in (4.12). If the underlying process is Gaussian, standard results for the multivariate Gaussian distribution show that (4.12) is the conditional mean of $z(\underline{x}_0)$ given $z(\underline{x}_1), ..., z(\underline{x}_n)$, so that it is globally optimal (without restriction to G_2) with respect to minimization of prediction mean-squared error. Since it is also the conditional mode it is also globally optimal with respect to minimization of (4.6).

The treatment given here has reproduced Matheron's result from a predictive estimating function point of view. Also it depends on the unrealistic assumptions that both the mean of the process and its covariance are known. Ordinary kriging and universal kriging extend the theory to parametric linear models for the mean and I consider a more general version of this in the next section, from a predictive estimating function point of view. However, the real advantages of this point of view (and scope for generalization) are, in my opinion, in the possibilities of extension to non-linear prediction. Such an extension is broadly analogous to Godambe and Kale's (1991) extended Gauss-Markov theory for parametric estimation and will be treated in detail elsewhere.

5 Prediction with Unknown Mean

Suppose that the mean function $\theta(\underline{x})$ is a known function, say, $\theta(\underline{x}; \underline{\beta})$ of a p-dimensional vector of unknown parameters $\underline{\beta}$. This includes the special cases of: (a) $\theta(\underline{x})$ an unknown but constant scalar, for which ordinary kriging has

been developed and (b) $\theta(\underline{x};\underline{\beta}) = \sum_{j=1}^{p} f_j(\underline{x})\beta_j$ corresponding to universal kriging. However, here, we allow the possibility that $\theta(\underline{x};\underline{\beta})$ is a non-linear function of $\underline{\beta}$. We retain the assumption that the covariance matrix is known but possibly a function of $\underline{\beta}, \Sigma_{ZZ}(\underline{\beta})$.

We consider joint estimation of $\underline{\beta}$, and prediction of $z(\underline{x}_0)$ at an unobserved location $\underline{x}_0$. The optimal estimating function for $\underline{\beta}$ is the quasi-score

$$D^T \Sigma_{ZZ}^{-1}(\underline{z}_n - \underline{\theta}_n(\underline{x};\underline{\beta})) \tag{5.1}$$

where D is the $n \times p$ matrix with (i,j) element $\frac{\partial\theta(\underline{x}_i;\underline{\beta})}{\partial\beta_j}$, $\underline{z}_n$ is the data vector as before and $\underline{\theta}_n(\underline{x};\underline{\beta})$, is the n-vector $(\theta(\underline{x}_1;\underline{\beta}),\ldots,\theta(\underline{x}_n;\underline{\beta}))$. The optimal prediction function for $z(\underline{x}_0)$, given $\underline{\beta}$, is given by (4.10) appropriately modified with $\theta_0 = \theta(\underline{x}_0;\underline{\beta})$, $\underline{\theta} = \underline{\theta}_n(\underline{x};\underline{\beta})$ and $\underline{d}$, Σ_{ZZ}, the specified known functions of $\underline{\beta}$. The optimal prediction function for $z(\underline{x}_0)$ can then be obtained by solving (5.1) for the maximum quasi-likelihood estimate, $\hat{\beta}_{QL}$, say, and inserting this into the modified version of (4.10). In general, solution of (5.1) will require a numerical solution. However, in the special case where $\theta(\underline{x};\underline{\beta}) = \sum_{j=1}^{p} f_j(\underline{x})\beta_j$, an explicit predictor can be found and shown to correspond to the universal kriging equations of Matheron. However, this derivation is more general in that it allows non-linear functions of $\underline{\beta}$ and additionally the covariance function may be a function of $\underline{\beta}$.

Finally, we remark that a partially Bayes approach similar to that of Godambe (1994) could be developed for spatial prediction. This involves the combination of estimating functions based on prior information about the mean function together with the quasi-score function based on the data. Although Godambe emphasizes parametric estimation he gives a brief illustration how this may be extended to forecasting of a future value of a branching process.

6 Optimal Prediction Functions Based on Likelihood

Whereas the Bayes approach to prediction is logically very appealing except for the sticking point of the prior, likelihood approaches are considerably murkier! Bjornstad (1990) gives a recent extensive survey outlining 14 different predictive likelihoods! This proliferation of definition suggests that predictive likelihood rests on somewhat shaky logical foundations. Bayarri, De Groot and Kadane (1987), in a provocative paper, question whether likelihood itself can be rigorously defined, in general, and much of their critique refers to situations in which prediction is of importance; see also, Berger and Wolpert (1988) for a discussion of this. In a sense, the entire parameter $\underline{\theta}$ is a nuisance parameter here so it is natural to consider elimination

methods for nuisance parameters analogous to conventional methods for the parameter of interest. Basically, the predictive likelihoods considered by Bjornstad fall into three categories: (i) elimination by profiling, (ii) elimination by conditioning on "sufficient" statistics for the (nuisance) parameter, (iii) elimination by integration. Method (i) entails the usual difficulties entailed in inserting MLEs for the nuisance parameter. Method (iii) is similar to the Bayesian approach and provides a probability distribution for the parameter (e.g. the Bayesian predictive density in section three is an instance of this). However, where a fiducial distribution is available for the parameter, Kalbfleisch (1971) shows how to obtain a predictive distribution for the future observation. This is related to Fisher's (1956) original argument. Method (ii) is a method for ordering the plausibility of future values proposed independently by Lauritzen (1974) and Hinkley (1979) and more generally by Butler (1986).

Another method, not discussed by Bjornstad, is that of marginal predictive likelihood discussed very briefly by Butler (1986). In this section we discuss this from the predictive estimation function point of view and conjecture that under certain model assumptions the marginal predictive score function has a certain optimality property. Let $\underline{x} = (x_1, ..., x_n)$ be the current data and z a scalar future observable. Suppose there exists a transformation

$$(\underline{x}, z) \to (T, A)$$

where (T, A) is sufficient for θ in the model $f_{\underline{x},z}(\underline{x}, z; \theta)$, $A = A(\underline{x}, z)$ is ancillary for θ, i.e. possesses a distribution not involving θ, and $T = T(\underline{x})$ is sufficient for θ in the model $f_{\underline{x}}(\underline{x}; \theta)$. Suppose also that the likelihood factorizes as

$$\begin{aligned} f_{\underline{x},z}(\underline{x}, z; \theta) &= f_{\underline{x},z|(T,A)}(\underline{x}, z) \\ &\quad f_A(a(\underline{x}, z)) f_{T|A}(T(\underline{x})|A; \theta). \end{aligned} \tag{6.1}$$

If, in addition, $T(\underline{x})$ is complete given A (conditionally complete) Basu's theorem (1959) implies that T and A are independent so that the final factor in (6.1) is the same as $f_T(T(\underline{x}); \theta)$.

The last two components of the above factorization in essence separate the "data" $(\underline{x}, z)$ into two components, $f_{T|A}(\cdot)$ which is relevant to inference about θ based on the sufficient statistics $T(\underline{x})$, and $f_A(\cdot)$ which provides information relevant to predicting z given $\underline{x}$. Classical frequentist prediction intervals often involve inversion of a function such as $A(\underline{x}, z)$ which is sometimes referred to as a predictive pivot. If the above factorization applied, it seems reasonable to consider the class of prediction functions which involve the data through A alone. We conjecture that the optimal prediction function in this class would correspond to the marginal predictive score function

based on differentiation with respect to z. A proof might be constructed motivated by arguments of Lloyd (1987). He considers estimation of a parameter of interest θ in the presence of a nuisance parameter ϕ and shows that the marginal likelihood based on a maximal ancillary for ϕ provides the optimal estimating function for θ, if the remainder of the likelihood is conditionally complete. In our case, the nuisance parameter is the entire parameter θ and the quantity of interest is z.

As a simple example of a situation to which the factorization applies consider $x_1, ..., x_n$ and z as independently distributed $N(\theta, 1)$, where z is to be predicted. Then (6.1) applies with $T(\underline{x}) = \overline{x}$ and $A(\underline{x}, z) = z - \overline{x}$.

An alternate factorization to (6.1), which is sometimes available, is when a statistic $T(\underline{x}, z)$, sufficient for θ, is available in the joint model for $\underline{x}$ and z. Predictive likelihoods which eliminate θ by conditioning on T can then be constructed along the lines of Hinkley (1979) or more generally Butler (1986). Such a factorization is available, for example, in exponential families. There is a parallel here with similar conditioning arguments for parametric estimation, where nuisance parameters can be eliminated by conditioning with respect to statistics sufficient for the nuisance parameter. In the predictive case, we note that the conditioning statistic T is a function of the future observation z, which corresponds to the "parameter" of interest in the estimative case. It is an open question whether optimal prediction functions based on conditional predictive likelihood can be constructed.

7 Discussion

This paper has considered the problem of predicting observables from a different perspective than more classical formulations. In the classical approach the two common desiderata are: (i) unbiasedness of the predictor, and (ii) minimization of prediction mean-squared error. As regards (i), the concept of unbiasedness is different from the usual unbiasedness concept for an estimator of a fixed parameter, as pointed out by Robinson (1991). The "ideal" predictor, if available, is the mean of the predictive distribution of the unobserved variable given those observed. Restrictions to linear unbiased predictors lead to best linear unbiased predictors, commonly referred to as BLUPs. By contrast, in the formulation given here, (i) above is replaced by: (i') unbiasedness of the prediction function and (ii) is replaced by: (ii') minimization of criterion (2.4).

Since (i') need not imply (i), biased predictors are possible in this formulation. Similarly (ii) and (ii') are equivalent only for a proper subclass of prediction functions so that the formulation via prediction functions is more general. The "ideal" predictor, if available, corresponds to the mode of the predictive distribution. If this is unavailable, projection into various sub-

spaces of prediction functions produces locally optimal prediction functions which are closest to the ideal prediction function in the L^2 norm defined by the covariance inner product defined on the space of prediction functions. The classical prediction theory has an analogous projection formulation but with a different inner product.

I have shown here that the new formulation reproduces classical prediction results in the subset of cases where they coincide. Godambe and Kale (1991), in the case of parametric estimation, show that the estimating function approach reproduces classical optimality results such as the Gauss-Markov theorem in elementary cases for the linear model, but offers a good deal more generality for non-linear models and quasi-likelihood models where the original Gauss-Markov approach fails. Those authors successfully develop an extended Gauss-Markov theory for these cases which is logically equivalent to the original Gauss-Markov theory for the elementary case. The development of analogous extensions for prediction functions will be treated in a separate publication. Areas of potential application being considered include non-linear geostatistical problems such as disjunctive kriging and transgaussian kriging.

Acknowledgement

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada, Grant No. A85584.

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis.* Cambridge: Cambridge University Press.

Bayarri, M. J., De Groot, M. H. and Kadane, J. B. (1987). What is the likelihood function? (with discussion). In *Statistical Decision Theory and Related Topics* IV, vol. 1, Gupta, S. S. and Berger, J., Eds., Springer-Verlag, New York.

Basu, D. (1959). The family of ancillary statistics. *Sankhya*, 21, 247-256.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle.* Hayward, California. Institute of Mathematical Statistics.

Bjornstad, J. F. (1990). Predictive likelihood: a review. *Statistical Science*, 5, 242-265.

Box, G. E. P. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Butler, R. W. (1986). Predictive likelihood inference with applications (with discussion). *J. Roy. Statist. Soc.*, B, 48, 1-38.

Butler, R. W. (1990). Comment on Bjornstad, J. F. Predictive likelihood: a review. *Statistical Science*, 5, 255-259.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22, 239-252.

Cressie, N. (1991). *Statistics for Spatial Data.* John Wiley and Sons, New York.

De Finetti, B. (1937). La prevision: ses lois logiques, ses sources subjectives. *Annales de l'Institute Henri Poincare*, 7, 1-68.

De Finetti, B. (1974, 1975). *Theory of Probability*, Volumes I and II. Wiley, New York.

Ferreira, P. (1982). Estimating equations in the presence of prior knowledge. *Biometrika*, 69, 667-669.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc., London, Ser. A*, 222, 309-368.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver and Boyd.

Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

Ghosh, M. (1990). On a Bayesian analog of the theory of estimating functions. In *C. G. Khatri Memorial Volume* of the Gujarat Statistical Review, 17A, 47-52.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31, 1208-1211.

Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika*, 72, 419-428.

Godambe, V. P. (1994). Linear Bayes and optimal estimation. Technical Report Series, Dept. of Statistics, University of Waterloo, Stat-94-11.

Godambe, V. P. and Kale, B. K. (1991). Estimating functions: an overview. In *"Estimating Functions"*, V. P. Godambe, Ed., Clarendon Press, Oxford.

Hinkley, D. V. (1979). Predictive likelihood. *Ann. Statist.*, 7, 718-728. Corrigendum 8, 694.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press, London.

Kalbfleisch, J. D. (1971). Likelihood methods of prediction. In *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott, Eds., Holt, Rinehart and Wilson, New York, 378-392.

Kolmogorov, A. N. (1941). Interpolation and extrapolation of stationary random sequences. *Izvestiia Akademii Nauk SSSR, Seriia Matematish-*

eskiia, 5, 3-14.

Lauritzen, S. L. (1974). Sufficiency, prediction and extreme models. *Scand. J. Statist.*, 1, 128-134.

Lindley, D. V. (1990). The 1988 Wald memorial lectures: The present position in Bayesian statistics (with discussion). *Statistical Science*, 5, 44-89.

Mathiasen, P. E. (1979). Prediction functions. *Scand. J. Statist.*, 6, 1-21.

Matheron, G. (1962). *Traite de Geostatistique Appliquee*, Tome 1, Memoires du Bureau de Recherche Geologiques et Minieres, No. 14, Editions Technip, Paris.

Matheron, G. (1976). A simple substitute for conditional expectation: The disjunctive kriging. *In Advanced Geostatistics in the Mining Industry*, M. Guarascio, M. David and C. Huijbregts, Eds., Reidel, Dordrecht, 221-236.

Naik-Nimbalkar, U. V. and Rajarshi, M. B. (1995). Filtering and smoothing via estimating functions. *J. American Stat. Assoc.*, 90, 301-306.

Pearson, K. (1920). The fundamental problem of practical statistics. *Biometrika* 13, 1-16.

Ripley, B. D. (1981). *Spatial Statistics.* Wiley, New York.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6, 15-51.

Stigler, S. M. (1976). Discussion of Savage, L. J., On rereading R. A. Fisher. *Ann. Stat.*, 4, 441-500.

Thavaneswaran, A. and Thompson, M. E. (1986). Optimal estimation for semi-martingales. *J. Appl. Prob.*, 23, 409-417.

Thavaneswaran, A. N. and Thompson, M. E. (1988). A criterion for filtering in semi-martingale models. *Stoch. Processes and their Applications*, 28, 259-265.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* MIT Press, Cambridge, MA.