A note on design-based versus model-based variance estimation in stereology

Asger Hobolth and Eva B. Vedel Jensen
Laboratory for Computational Stochastics
University of Aarhus

Abstract

Recently, systematic sampling on the circle and the sphere has been studied by Gual-Arnau and Cruz-Orive (2000) from a design-based point of view. In this note, it is shown that their mathematical model for the covariogram is in a model-based statistical setting a special case of the p-order shape model, suggested in Hobolth *et al.* (1999, 2000) for planar objects without landmarks. Benefits of this observation include an alternative variance estimator, applicable in the original problem of systematic sampling. In a wider perspective, the paper contributes to the discussion concerning design-based versus model-based stereology.

Keywords: covariogram, circulant matrix, Fourier series, planar objects, shape, stationarity, stereology, systematic sampling.

1 Introduction

In stereology, the aim is typically to make inference about a population of spatial objects from geometric samples of the objects such as line and plane sections. The objective is not to reconstruct the objects, but instead to make inference about quantitative properties such as volume or surface area. If a typical object from the population can be regarded as a realization of a stochastic process R, then the quantitative property of interest can be expressed as a function f of R. Using a geometric sampling design ϕ, independent of R, a predictor $\hat{f}(R,\phi)$ of $f(R)$ can often be constructed, based on reasoning from stochastic geometry, which is design-unbiased, i.e.

$$\mathrm{E}(\hat{f}(R,\phi)|R) = f(R).$$

It is part of the methodology of design-based stereology to construct a design-unbiased estimator $\hat{\sigma}^2_R(\phi)$ of the conditional variance

$$\sigma^2_R = \mathrm{Var}(\hat{f}(R,\phi)|R).$$

The estimator $\hat{\sigma}^2_R(\phi)$ thus satisfies

$$\mathrm{E}(\hat{\sigma}^2_R(\phi)|R) = \sigma^2_R.$$

Usually $\hat{\sigma}^2_R(\phi)$ is based on the empirical covariogram.

In most cases it is of interest to make statements about the population of objects and not only about the sampled objects. A relevant quantity is here the prediction error

$$\mathrm{E}(\hat{f}(R,\phi) - f(R))^2.$$

Using that $\hat{f}(R,\phi)$ and $\hat{\sigma}^2_R(\phi)$ are design-unbiased, the prediction error can be rewritten as

$$\begin{aligned} & \mathrm{E}(\hat{f}(R,\phi) - f(R))^2 \qquad (1.1) \\ & = \mathrm{Var}\big(\hat{f}(R,\phi) - f(R)\big) \\ & = \mathrm{Var}\big(\mathrm{E}(\hat{f}(R,\phi) - f(R)|R)\big) + \mathrm{E}\big(\mathrm{Var}(\hat{f}(R,\phi) - f(R)|R)\big) \\ & = \mathrm{E}\big(\mathrm{Var}(\hat{f}(R,\phi)|R)\big) \\ & = \mathrm{E}\sigma^2_R = \mathrm{E}\hat{\sigma}^2_R(\phi). \qquad (1.2) \end{aligned}$$

Therefore, $\hat{\sigma}^2_R(\phi)$ or an average of such estimators for a sample of objects can be regarded as an unbiased estimator of the prediction error.

In the present paper we propose the alternative of using a likelihood-based method of estimating the prediction error. The discussion is centred around the example where $R = \{R(2\pi t) \in \mathbb{R} : 0 \leq t \leq 1\}$ is a 2π periodic stochastic process and

$$f(R) = \int_0^1 R(2\pi t)dt$$

is the quantity of interest. Based on $n \geq 2$ equally spaced measurements

$$\{R(2\pi(\phi + j/n)) : j = 0, \ldots, n-1\}$$

of the stochastic process R, with ϕ uniformly distributed in $[0, 1/n]$, Gual-Arnau and Cruz-Orive (2000) have recently suggested a design-unbiased estimator of the conditional variance of

$$\hat{f}(R,\phi) = \frac{1}{n}\sum_{j=0}^{n-1} R(2\pi(\phi + j/n)). \qquad (1.3)$$

In this paper we suggest a parametric model for R with a covariance structure similar to that of Gual-Arnau and Cruz-Orive (2000). The prediction error is estimated by inserting the maximum likelihood estimates of the model parameters into a closed form parametric expression for the prediction error. The proposed estimator of the prediction error is optimal under the suggested model for R.

The paper is organised as follows. In Section 2 we recall the design-based variance estimation of Gual-Arnau and Cruz-Orive (2000), while the likelihood-based variance estimation is carried out in Section 3. In Section 4 it is shown that the proposed model is a special case of the p-order shape model, suggested in Hobolth *et al.* (1999, 2000) for planar objects without landmarks. It is also pointed out that a similar discussion about estimation procedures has taken place in the geostatistical community during the last decade.

2 Design-based variance estimation

Let $R = \{R(2\pi t) \in \mathbb{R} : 0 \leq t \leq 1\}$ be a 2π periodic stochastic process, which is of bounded variation, square integrable and piecewise continuous, and let $\phi \sim \mathrm{U}[0, 1/n]$

be independent of R. If we define $\hat{f}(R,\phi)$ as in (1.3), then Cruz-Orive and Gual-Arnau (2000) treat the problem of estimating the conditional variance $\mathrm{Var}(\hat{f}(R,\phi)|R=r)$. In particular they show that, cf. Gual-Arnau and Cruz-Orive (2000, Corollary 2.1),

$$\mathrm{Var}(\hat{f}(R,\phi)|R=r) = \sum_{k\in\mathbb{Z}\setminus\{0\}} c_{kn}, \tag{2.1}$$

where

$$c_k = \int_0^1 g(t)e^{-2\pi ikt}dt,\ k\in\mathbb{Z},$$

are the Fourier coefficients of the covariogram

$$g(t) = \int_0^1 r(2\pi h)r(2\pi(h+t))dh,\ 0\le t\le 1.$$

Here and throughout the paper we use periodic extensions of the functions (i.e. $r(2\pi(x+k)) = r(2\pi x),\ k\in\mathbb{Z}$). Note that c_k is real and $c_k = c_{-k}$ because $g(1-t) = g(t)$. The covariogram

$$g(t) = \sum_{k\in\mathbb{Z}} c_k e^{2\pi ikt} = c_0 + 2\sum_{k=1}^{\infty} c_k\cos(2\pi kt), \tag{2.2}$$

is modelled by a polynomium of order $2p,\ p\in\mathbb{N}$. The fact that $g(t) = g(1-t)$ causes restrictions on the coefficients of the polynomium. Gual-Arnau and Cruz-Orive (2000, p. 635) show that in fact the polynomium only depends on two real parameters β_0, β, and that the Fourier coefficients of g take the form

$$c_0 = \beta_0 - \sum_{k\in\mathbb{Z}\setminus\{0\}} c_k,\ \ c_k = \frac{(2p)!}{k^{2p}}\beta,\ k\in\mathbb{Z}\setminus\{0\}, \tag{2.3}$$

where $c_0, \beta > 0$. Unbiased estimators of $g(0)$ and $g(1/n)$ are obtained by

$$\begin{aligned}\hat{g}(0) &= \frac{1}{n}\sum_{j=0}^{n-1} r(2\pi(\phi+j/n))^2,\\ \hat{g}(1/n) &= \frac{1}{n}\sum_{j=0}^{n-1} r(2\pi(\phi+j/n))r(2\pi(\phi+(j+1)/n)),\end{aligned}$$

and using the formula for the Bernoulli polynomium, cf. e.g. Abramovitz and Stegun (1970, p. 805),

$$B_{2p}(t) = \frac{(-1)^{p-1}(2p)!}{(2\pi)^{2p}} 2\sum_{k=1}^{\infty}\frac{\cos(2\pi kt)}{k^{2p}},\ \ 0\le t\le 1,\ \ p\in\mathbb{N},$$

an unbiased estimator of $\mathrm{Var}(\hat{f}(R,\phi)|R=r)$ given by

$$\frac{\hat{g}(0)-\hat{g}(2\pi/n)}{n^{2p}}\frac{1}{1-B_{2p}(1/n)/B_{2p}}, \tag{2.4}$$

is obtained, where $B_{2p}=B_{2p}(0)$ is the Bernoulli number of order $2p$. Note that

$$2n(\hat{g}(0)-\hat{g}(2\pi/n))=\sum_{j=0}^{n-1}\left(r(2\pi(\phi+j/n))-r(2\pi(\phi+(j+1)/n)\right)^2,$$

and thus the estimator is based on first-order differences.

3 Model-based setting

Now we recast the models and estimation procedures of Gual-Arnau and Cruz-Orive (2000) in terms of a stationary, random periodic process R with mean μ and covariance function

$$\sigma(t)=\sum_{k\in\mathbb{Z}}\lambda_k e^{2\pi ikt}=\lambda_0+2\sum_{k=1}^{\infty}\lambda_k\cos(2\pi kt),\ \ 0\le t\le 1.$$

Note that the λ_k's are real because $\sigma(1-t)=\sigma(t)$. If we make a Fourier expansion of the random covariogram

$$G(t)=\int_0^1 R(2\pi h)R(2\pi(h+t))dh=C_0+2\sum_{k=1}^{\infty}C_k\cos(2\pi kt)$$

then $EC_k=\lambda_k,\ k\ge 1$, and $EC_0=\lambda_0+\mu^2$. Accordingly, the covariogram model (2.3) corresponds to a covariance function $\sigma(t)$ with

$$\lambda_0=\beta_0-\sum_{k\in\mathbb{Z}\setminus\{0\}}\lambda_k,\ \lambda_k=\frac{(2p)!}{k^{2p}}\beta,\ k\in\mathbb{Z}\setminus\{0\}. \tag{3.1}$$

Note that

$$\sigma(0)=\lambda_0+\sum_{k\in\mathbb{Z}\setminus\{0\}}\lambda_k=\beta_0,$$

which means that β_0 determines the variance and β/β_0 the correlation structure.

In a model-based setting, the aim is to estimate the error involved in using

$$\hat{f}(R,\phi)=\frac{1}{n}\sum_{j=0}^{n-1}R(2\pi(\phi+j/n))$$

as a predictor of $f(R) = \int_0^1 R(2\pi t)dt$. In terms of model parameters the prediction error is given by, cf. (1.1), (2.1) and (3.1),

$$\begin{aligned} \mathrm{E}(\hat{f}(R,\phi) - f(R))^2 &= \mathrm{E}(\mathrm{Var}(\hat{f}(R,\phi)|R)) \\ &= \sum_{k\in\mathbb{Z}\setminus\{0\}} \lambda_{kn} \\ &= \beta \sum_{k\in\mathbb{Z}\setminus\{0\}} \frac{(2p)!}{(kn)^{2p}} \\ &= (-1)^{p-1}(2\pi)^{2p} B_{2p} \frac{1}{n^{2p}} \beta. \end{aligned} \tag{3.2}$$

Note that the prediction error will be the same if we fix $\phi = 0$, say.

We can use the procedure suggested by Gual-Arnau and Cruz-Orive (2000) for obtaining an unbiased estimator of the prediction error in the model-based setting. Another approach is to estimate β in the parametric model for R by maximum likelihood estimation. Suppose for instance that the process R is Gaussian. Then the vector

$$R_n = \big(R(2\pi\phi), R(2\pi(\phi + 1/n)), \ldots, R(2\pi(\phi + (n-1)/n))\big)^T$$

follows a multivariate normal distribution with mean $(\mu, \ldots, \mu)^T = \mu 1_n^T$ and an $n \times n$ circulant covariance matrix

$$\Sigma = \mathrm{circ}\big(\sigma(0), \sigma(1/n), \ldots, \sigma((n-1)/n)\big).$$

The covariance matrix can be diagonalised by the complex $n \times n$ discrete Fourier transform matrix W with entries $w_{jk} = e^{2\pi ijk/n}/n$, $0 \le j, k \le n-1$, cf. e.g. Wei (1990, Chapter 10). Let w_k denote the $(k+1)$'th coloumn of W so that $W = [w_0, \ldots, w_{n-1}]$ and let $W^* = \overline{W}^T$ denote the complex conjugate of W. Then

$$W^*\Sigma W = \mathrm{diag}(\tilde{\lambda}_0, \ldots, \tilde{\lambda}_{n-1})$$

is a diagonal matrix with

$$\tilde{\lambda}_j = w_j^* \Sigma w_j = \sum_{k\in\mathbb{Z}} \lambda_{j+nk}, \quad j = 0, \ldots, n-1,$$

on the diagonal. Note that only the parameter β is present in the expression of $\tilde{\lambda}_j, j = 1, \ldots, n-1$, while both β and β_0 are present in the expression of $\tilde{\lambda}_0$. Similarly we find that

$$\begin{aligned} (R_n - \mu 1_n)^* \Sigma^{-1} (R_n - \mu 1_n) &= (R_n - \mu 1_n)^* n W W^* \Sigma^{-1} n W W^* (R_n - \mu 1_n) \\ &= \frac{(\hat{f} - \mu)^2}{\tilde{\lambda}_0} + \sum_{j=1}^{n-1} \frac{\hat{\lambda}_j}{\tilde{\lambda}_j} \\ &= \frac{(\hat{f} - \mu)^2}{\tilde{\lambda}_0} + \beta \sum_{j=1}^{n-1} \frac{\hat{\lambda}_j}{\tilde{\kappa}_j}, \end{aligned}$$

where

$$\hat{f} = \hat{f}(R, \phi), \quad \hat{\lambda}_j = \hat{\lambda}_j(R, \phi) = w_j^* R_n R_n^* w_j, \ j = 0, \ldots, n-1,$$

and

$$\tilde{\kappa}_j = \tilde{\lambda}_j / \beta = \sum_{k \in \mathbb{Z}} \frac{(2p)!}{(j+nk)^{2p}}, \ j = 1, \ldots, n-1. \tag{3.3}$$

Thus the sufficient statistic is given by

$$T = (\hat{f}, \hat{f}^2, \sum_{j=1}^{n-1} \hat{\lambda}_j / \tilde{\kappa}_j).$$

According to the Rao-Blackwell theorem any function of T is the minimum variance estimator of its mean value. Furthermore it follows from the theory of exponential families that T is complete, and hence

$$\hat{\beta} = \frac{1}{n-1} \sum_{j=1}^{n-1} \frac{\hat{\lambda}_j}{\tilde{\kappa}_j}$$

is the unique unbiased estimator of β with minimum variance. Using the real discrete Fourier transform matrix similar calculations show that $\hat{\beta}$ follows a $\beta\chi^2(n-1)/(n-1)$ distribution, and therefore we can supply the point estimate of β with a confidence interval. This is an important option which does not exist in a design-based setting.

The likelihood-based estimator of β is a weighted sum of the squared length of the discrete complex Fourier coefficients $\hat{\lambda}_j$. It is clear from (3.3) that the weights $\tilde{\kappa}_j$ depend crucially on the order $2p$ of the polynomium. Below we discuss how p relates to the smoothness of the sample paths, and may be considered as a third parameter in the model.

We estimate the prediction error by, cf. (3.2),

$$(-1)^{p-1}(2\pi)^{2p} B_{2p} \frac{1}{n^{2p}} \hat{\beta}. \tag{3.4}$$

It is worth noticing that for $n = 2$ and $n = 3$ this estimator actually coincides with the estimator (2.4) of Gual-Arnau and Cruz-Orive (2000). Note also that in a design-based setting, (3.4) is an unbiased estimator of $\mathrm{Var}(\hat{f}(R, \phi)|R = r)$ under the covariogram model (2.3).

4 Discussion

4.1 The p-order shape model

The model (3.1) is a special case of the p-order model suggested in Hobolth *et al.* (1999, 2000) which appears to be very natural for modelling the shape of planar

objects K without landmarks. In this setting K is assumed to be star-shaped with respect to a fixed point $z \in K$, and $R(2\pi t)$ is the radius-vector function evaluated at $2\pi t$, i.e. the distance from z to the boundary of K along a line with angle $2\pi t$ relative to a fixed axis. In the p-order model the λ_k's are determined by

$$\lambda_0 \geq 0, \quad \lambda_k^{-1} = \tilde{\alpha} + \tilde{\beta} k^{2p}, \quad k \in \mathbb{Z} \backslash \{0\}, \tag{4.1}$$

where $\tilde{\alpha} \geq 0$, $\tilde{\beta} > 0$, $p > 1/2$. Note that in this model p is a parameter and not a fixed integer as in (2.3). For $\tilde{\alpha} = 0$ and $1/\tilde{\beta} = (2p)!\beta$ we get the model (3.1). In Hobolth *et al.* (2000) it is discussed how the parameters $(\tilde{\alpha}, \tilde{\beta}, p)$ relates to the shape of the object. The parameter p determines the smoothness of the object boundary. In the Gaussian case the sample paths are k times continuously differentiable, where k is the integer satisfying $p \in]k - 1/2, k + 1/2[$. For fixed p, $\tilde{\alpha}$ determines the global shape while $\tilde{\beta}$ determines the local shape. Furthermore, it can be argued that λ_1 relates to asymmetry of K relative to $z \in K$, so the regression model (4.1) should for geometrical reasons only be considered for $|k| \geq 2$. In Hobolth *et al.* (2000) it is demonstrated how the three parameters can be estimated using maximum likelihood. Based on the observed information it is also possible to determine confidence intervals of the parameters.

In geometric examples, R is typically a power (2 or 3) of the radius-vector function. In such cases, a Gaussian assumption may not be appropriate. Hobolth *et al.* (2000) provide tools for analysing non-Gaussian processes in this context.

4.2 Covariogram versus likelihood-based methods

The estimation procedure of Gual-Arnau and Cruz-Orive (2000) is based on the empirical covariogram, while we suggest a likelihood-based method. In the geostatistical community a discussion of the two procedures have taken place during the last decade, and has resulted in a move towards the adoption of likelihood-based methods (Diggle *et al.*, 1998, p. 305). We refer the interested reader to the recent monograph Stein (1999) and references therein for more information on parameter estimation using covariogram- or likelihood-based methods. We believe that a corresponding discussion is needed among the stereologists and we hope with this paper to have contributed in a constructive manner to such a discussion.

Acknowledgement

We are very grateful to Jan Pedersen for his valuable comments on this work. This work was supported in part by MaPhySto, funded by a grant from the Danish National Research Foundation.

References

Abramovitz, M. and Stegun, I.A. (1965). *Handbook of Mathematical Functions.* Dover, New York.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Appl. Statist.* **47**, 299-350.

Gual-Arnau, X. and Cruz-Orive, L.M. (2000). Systematic sampling on the circle and on the sphere. *Adv. Appl. Prob. (SGSA)* **32**, 628-647.

Hobolth, A., Kent, J.T. and Dryden, I.L. (1999). On the relation between edge and vertex modelling. *Research Report* **7**, Laboratory for Computational Stochastics, University of Aarhus. To appear in *Scand. J. Statist.*

Hobolth, A., Pedersen, J. and Jensen, E.B.V. (2000). A continuous parametric shape model. *Research Report* **13**, Laboratory for Computational Stochastics, University of Aarhus. Submitted.

Stein, M.L. (1999). *Interpolation of Spatial Data.* Springer, New York.

Wei, W.W.S (1990). *Time Series Analysis.* Addison-Wesley, Redwood City, California.