THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

On variance estimation in circular systematic sampling

Kristjana Ýr Jónsdóttir, Eva B. Vedel Jensen
and Asger Hobolth

Research Report

No. 06 | June 2005

On variance estimation in circular systematic sampling

This Thiele Research Report is also Research Report number 459 in the Stochastics Series at Department of Mathematical Sciences, University of Aarhus, Denmark.

On variance estimation in circular systematic sampling [1]

Kristjana Ýr Jónsdóttir, Eva B. Vedel Jensen and Asger Hobolth
University of Aarhus

Abstract

An extended covariogram model is proposed for estimating the precision of circular systematic sampling. The extension is motivated by recent developments in shape analysis of featureless planar objects.

1. Introduction

Recently, the precision of systematic sampling on the circle has been discussed in Gual-Arnau and Cruz-Orive (2000) and Cruz-Orive and Gual-Arnau (2002). In particular, variance prediction formulae based on a global polynomial model for the covariogram have been developed. In Hobolth and Jensen (2002), this approach is discussed both in a design-based and a model-based setting, and an alternative model-based method of estimating the parameter of the covariogram is described.

In this note, we summarize these developments and argue that it may be natural to consider an extension of the polynomial covariogram model, see also the discussion in Hobolth and Jensen (2002). We explain the geometric interpretation of the parameters of the proposed extended model and report preliminary simulation results.

2. A global polynomial covariogram model

In a design-based setting, the parameter Q to be estimated is of the form

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

where $f : [0, 2\pi] \to \mathbb{R}_+$ is of bounded variation, square integrable and piecewise continuous. An unbiased estimate of Q, based on a systematic sample of fixed size n, is

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

where ϕ is uniform random in the interval $[0, \frac{1}{n}[$. The covariogram of f is given by

$$g(t) = \int_0^1 f(2\pi h)f(2\pi(h+t))dh,$$

[1] This paper is a contribution to the special issue of Journal of Microscopy to be published from the "Thiele Workshop on Variance Estimation in Stereology", 22 - 26 November 2004, University of Aarhus, Denmark.

$0 \leq t < 1$, where we use a periodic extension of f. The covariogram satisfies $g(1-t) = g(t)$ and therefore its Fourier expansion takes the form

$$g(t) = d_0 + 2\sum_{k=1}^{\infty} d_k \cos(2\pi k t), \tag{1}$$

$0 \leq t < 1$, where the Fourier coefficients are given by

$$d_k = \int_0^1 g(t)\cos(2\pi k t)dt, \quad k = 0, 1, \ldots .$$

Moreover, it can be shown that

$$\mathrm{Var}(\hat{Q}(f, \phi, n)) = 2\sum_{k=1}^{\infty} d_{kn}. \tag{2}$$

Gual-Arnau and Cruz-Orive (2000) suggest to estimate $\mathrm{Var}(\hat{Q}(f, \phi, n))$ by modelling the covariogram g by a polynomial of order $2p$, $p \in \mathbb{N}$, of the following form

$$g(t) = \sum_{j=0}^{p} \beta_{2j} t^{2j} + \beta_{2p-1} t^{2p-1}, \tag{3}$$

where $0 \leq t < 1$, $p \in \mathbb{N}$ and β_j are unknown parameters. In Gual-Arnau and Cruz-Orive (2000) it is shown that since $g(t) = g(1-t)$, the polynomial in (3) depends only on two real parameters β_0 and β and is given by

$$g(t) = \beta_0 + (-1)^p (2\pi)^{2p} (B_{2p} - B_{2p}(t))\beta, \tag{4}$$

$0 \leq t < 1$, where $B_{2p}(t)$ is a Bernoulli polynomial of order $2p$ and $B_{2p} = B_{2p}(0)$. For more details on Bernoulli polynomials, see Abramovitz and Stegun (1965). Equivalently, the Fourier coefficients of g are on the form

$$\begin{aligned} d_0 &= \beta_0 - 2\sum_{k=1}^{\infty} d_k, \\ d_k &= \frac{(2p)!}{k^{2p}}\beta, \quad k = 1, 2, \ldots . \end{aligned} \tag{5}$$

Using (2) the variance can be written as

$$\begin{aligned} &\mathrm{Var}(\hat{Q}(f, \phi, n)) \\ &\quad = \frac{1}{n^{2p}}(-1)^{p-1}(2\pi)^{2p} B_{2p}\beta, \\ &\quad = \frac{1}{n^{2p}} \frac{g(0) - g(t)}{1 - B_{2p}(t)/B_{2p}}, \qquad 0 \leq t < 1. \end{aligned} \tag{6}$$

In Gual-Arnau and Cruz-Orive (2000), the following unbiased estimate of $\mathrm{Var}(\hat{Q}(f, \phi, n))$ is suggested

$$\frac{1}{n^{2p}} \frac{\hat{g}(0) - \hat{g}(\frac{1}{n})}{1 - B_{2p}(\frac{1}{n})/B_{2p}}, \tag{7}$$

where

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

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

are unbiased estimators of $g(0)$ and $g(\frac{1}{n})$, respectively. Note that this estimator only uses the empirical covariogram $\hat{g}$ near the origin. In Cruz-Orive and Gual-Arnau (2002), they suggest a more general estimator

$$\frac{1}{n^{2p}}\frac{1}{[\frac{n}{2}]}\sum_{k=1}^{[\frac{n}{2}]}\frac{\hat{g}(0)-\hat{g}(\frac{k}{n})}{1-B_{2p}(\frac{k}{n})/B_{2p}},$$

using more values of the empirical covariogram.

The global covariogram model can be formulated in a model-based framework by assuming that the function f is a realization of a stationary, random periodic stochastic process

$$F = \{F(2\pi t) : 0 \le t < 1\}$$

with mean μ and covariance function

$$\sigma(t) = \lambda_0 + 2\sum_{k=1}^{\infty}\lambda_k \cos(2\pi k t),$$

$0 \le t < 1$. The global covariogram model of Gual-Arnau and Cruz-Orive (2000), see (1) and (5), corresponds to assuming that

$$\lambda_0 = \beta_0 - 2\sum_{k=1}^{\infty}\lambda_k,$$

$$\lambda_k = \frac{(2p)!}{k^{2p}}\beta, \quad k = 1, 2, \ldots, \tag{8}$$

in the model-based setup. In Hobolth and Jensen (2002), it is shown that the prediction error of using $\hat{Q}(F,\phi,n)$ as an estimate of

$$Q = \int_0^1 F(2\pi t)dt$$

is given by

$$\begin{aligned}\mathbb{E}(\hat{Q}(F,\phi,n)-Q)^2 \\ &= 2\sum_{k=1}^{\infty}\lambda_{kn} \\ &= \frac{1}{n^{2p}}(-1)^{p-1}(2\pi)^{2p}B_{2p}\beta. \end{aligned}\tag{9}$$

Note that the prediction error is of the same form as the variance (6) in the design-based setup. Hobolth and Jensen (2002) suggest to estimate the parameter β using maximum likelihood estimation instead of using the empirical covariogram. It is shown that if F is assumed to be a Gaussian process, there exists a unique unbiased estimator of β with minimum variance. If we have n systematic observations of F,

$$F_n = \Big(F\Big(2\pi(\phi + \frac{j}{n}\Big)\Big) : j = 0, 1, \dots, n-1)^T,$$

then

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

where

$$\begin{aligned}
\hat{\lambda}_j &= \omega_j^* F_n F_n^* \omega_j, & j &= 0, \dots, n-1\\
\tilde{\kappa}_j &= \sum_{k\in\mathbb{Z}} \frac{(2p)!}{(j+nk)^{2p}}, & j &= 1, \dots, n-1\\
\omega_{j,k} &= \tfrac{1}{n} e^{2\pi i jk/n}, & j, k &= 0, \dots, n-1.
\end{aligned}$$

Here $\omega^* = \overline{\omega}^T$ denotes the complex conjugate of ω.

Using this maximum likelihood estimate of β we can estimate the prediction error by

$$(-1)^{p-1}(2\pi)^{2p} B_{2p}\frac{1}{n^{2p}}\hat{\beta}, \tag{10}$$

and it can be shown that for $n = 2$ and $n = 3$ this estimator coincides with the estimator (7) in the design-based setup.

3. An extension of the global covariogram model

The covariogram model (5) and its model-based analogue (8) can be motivated by tractability. It turns out, however, that it is natural from a geometric point of view to consider an extension of this model. In a model-based setting, the extended model is known as the $p-$order model, cf. Hobolth et al. (2002) and Hobolth et al. (2003). The covariance function of the extended model is determined by Fourier coefficients of the form

$$\lambda_0 \geq 0, \quad \lambda_k^{-1} = \tilde{\alpha} + \tilde{\beta}(k)^{2p}, \quad k = 1, 2 \dots, \tag{11}$$

where $\tilde{\alpha} \geq 0$, $\tilde{\beta} > 0$ and $p > \frac{1}{2}$. It can be shown that p determines the smoothness of the stochastic process F. In fact, if we assume that F is a Gaussian process, F is $k-1$ times continuously differentiable where k is the integer satisfying $p \in]k - \frac{1}{2}, k + \frac{1}{2}]$. For fixed p, $\tilde{\alpha}$ and $\tilde{\beta}$ determines the global and local fluctuations of the stochastic process F, respectively. Small values of $\tilde{\alpha}$ provide large fluctuations of the process on a global scale, while large values give smaller fluctuations. Also, the smaller $\tilde{\beta}$, the more fluctuations of F on a local scale.

In particular, if $F = R$, where R is the radial function of a planar object K, star-shaped relative to $z \in K$,

$$R(2\pi t) = \max\{r : z + r(\cos(2\pi t), \sin(2\pi t)) \in K\},$$

$0 \leq t < 1$, then p determines the smoothness of the boundary of the object K and for fixed p, $\tilde{\alpha}$ and $\tilde{\beta}$ determine the global and local shape of the object, respectively. The smaller $\tilde{\alpha}$, the more deviations from a circular shape of K are expected. Typically, in addition, the parameter λ_1 is set to zero if the point z is approximately the center of mass of the object K. For more details see Hobolth et al. (2003).

As mentioned above the model described in (8) is a special case of the p-order model, corresponding to

$$p \in \mathbb{N}, \tilde{\alpha} = 0, \tilde{\beta} = ((2p)!\beta)^{-1}.$$

Since $\tilde{\alpha} = 0$, this would provide large fluctuations of the stochastic process F on a global scale. If the geometric quantity of interest is the area, then $F \propto R^2$, and one would expect an object with large deviations from a circular shape. As a consequence, it seems natural to include an additional parameter $\tilde{\alpha}$ to allow for more flexibility. A detailed demonstration of how to estimate the parameters in the p-order model can be found in Hobolth et al. (2003).

In the design-based setup it seems also more natural to include an additional parameter α in the global covariogram model. Instead of modelling the Fourier coefficients of the covariogram by (5) we let

$$d_0 \geq 0, \quad d_k^{-1} = \tilde{\alpha} + \tilde{\beta} k^{2p}, \quad k = 1, 2, \ldots,$$

where $\tilde{\alpha} \geq 0$, $\tilde{\beta} > 0$ and $p > \frac{1}{2}$.

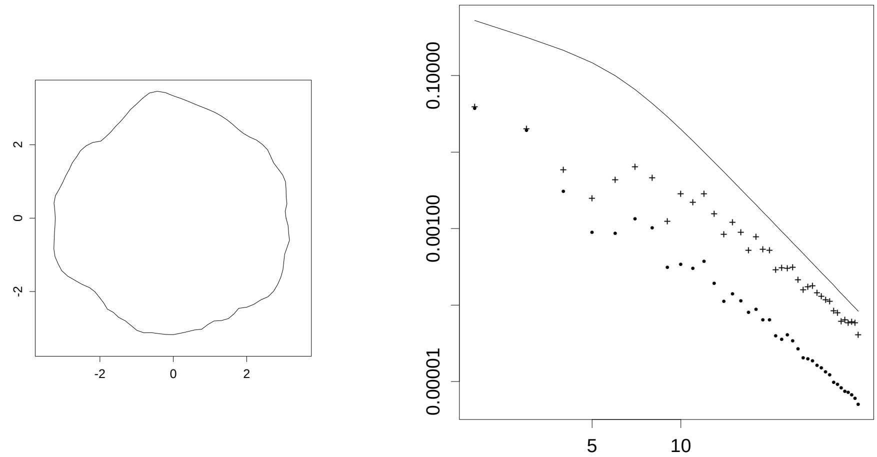

Figure 1: Left: The object K. Right: A log-log plot of the true prediction error (solid curve) as a function of n, together with the estimated prediction error (10), shown as $+$, and the variance estimate (7), shown as $\bullet$.

As an illustration, Figure 1, left, shows an object K with radial function R. Note that this object has a close to circular shape, so the process $F = R^2$ is not expected to have large global fluctuations. In fact, this object is the result of a simulation of the square root of a Gaussian 2nd order model with parameters $\mu = 10$, $\tilde{\alpha} = 5$ and $\tilde{\beta} = 0.005$. Figure 1, right, shows in a log-log plot the true prediction error

$$\mathbb{E}(\hat{Q}(F, \phi, n) - Q)^2 = 2\sum_{k=1}^{\infty} \lambda_{kn}$$

as a function of n (solid curve), the estimated prediction error (10) as a function of n (+) under the simplified model (8) with $\tilde{\alpha} = 0$ and the variance estimate (7) as a function of n (•). As expected, the prediction of the variance is not very good and this holds for both small and large n. The variance estimate (7) appears still to be of order n^{-2p}, but the level is too low. Note also that if $\tilde{\alpha} = 0$, the solid curve would be linear with slope $2p$. Clearly this is not the case for small n.

These results call for a closer investigation of the extended model and its use in assessment of the precision of circular systematic sampling.

References

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

Cruz-Orive, L. M. and Gual-Arnau, X. (2002). Precision of circular systematic sampling. *J. Microsc.*, 207:225–242.

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. and Jensen, E. B. V. (2002). A note on design-based versus model-based variance estimation in stereology. *Adv. Appl. Prob. (SGSA)*, 34:484–490.

Hobolth, A., Kent, J. T., and Dryden, I. L. (2002). On the relation between edge and vertex modelling in shape analysis. *Scand. J. Statist.*, 29:355–374.

Hobolth, A., Pedersen, J., and Jensen, E. B. V. (2003). A continuous parametric shape model. *Ann. Inst. Statist. Math.*, 55:227–242.