The Variogram Estimators of the Stationary Stochastic Processes

Tsekhavaya Tatsiana

Belarusian State University
Nezavisimosti Ave., 4, 220030, Minsk, Republic of Belarus
Tsekhavaya@bsu.by

Abstract. *The development of methods to detect investigation of an intrinsically-stationary stochastic processes and fields is a very important problem in data analysis. Variogram is a main characteristic intrinsically-stationary stochastic processes in time area. It is used for measuring the variability in space. The usual estimators of variogram are highly non-robust. We can find several alternative (robust) examples for estimation in the literature. The theory of construction of robust variogram estimators of the intrinsically-stationary random processes is developed in this paper. The estimators of variogram are constructed and its statistical properties are investigated.*

1 Introduction

Consider a random process $X(s), s \in R$. Then intrinsic stationarity is defined through first differences:

$$M\left\{X(s_1) - X(s_2)\right\} = 0, \quad D\left\{X(s_1) - X(s_2)\right\} = 2\gamma(s_1 - s_2),$$

$s_1, s_2 \in R$ [1]. The quantity $2\gamma(s)$ is known as the variogram. The variogram depends only on the relative position of the two variables $X(s_1)$ and $X(s_2)$. The semi-variogram $\gamma(s)$ is often synonymously named variogram when the usage is clear.

The term of variogram was coined by G. Matheron [2], although earlier appearances of this function can be found in the scientific literature. The variogram is the most important tool for the structural analysis of spatial phenomenons. It is the basis for further analysis, for example interpolation or extrapolation of values including the precision of estimation.

2 Estimators of the Variogram.

It is a matter of common experience that values often do not follow the normal distributions assumed for them, but, instead, follow some other heavier-tailed distribution. In this article we discuss the robust estimation of the variogram when the distribution is normal-like in the central region but heavier than normal in the tails. It is shown that the use of M-estimation yields stable robust estimates of the variogram.

Let $X(s), s \in R$, be a zero-mean intrinsically-stationary stochastic process of independent increments, with unknown variogram $2\gamma(s), s \in R$. Define

$$U_s(h) = (X(s+h) - X(s))^2, s, h \in R.$$

Then $MU_s(h) = 2\gamma(h)$. We can write $2\gamma(h)$ in the following way:

$$2\gamma(h) = M\sum_{i=0}^{h-1}\{X(s+h-i) - X(s+h-i-1)\}^2 = h2\gamma(1).$$

Therefore, the estimation of the variogram $2\gamma(1)$ becomes a problem of estimating the expectation of random of $U_s(1), s \in R$.

The estimators of variogram $2\gamma(1)$ in terms of sequence of observations

$$X(1), X(2), ..., X(n)$$

are defined as:

1. The mean

$$\bar{U}(1) = \frac{1}{n-1}\sum_{s=1}^{n-1} U_s(1),$$

where $U_s(1) = (X(s+1) - X(s))^2$;

2. The median

$$\tilde{U}(1) = \begin{cases} \overline{U}_{\frac{n}{2}}(1), & (n-1) = 2m+1, \\ 0,5\Big(\overline{U}_{\frac{n-1}{2}}(1) + \overline{U}_{\frac{n-1}{2}+1}(1)\Big), & (n-1) = 2m, \end{cases}$$

where $\overline{U}_1(1) \le ... \le \overline{U}_{n-1}(1)$ represent the ordered sample $U_s(1), s = \overline{1, n-1}$, $m = 1, 2, ...$;

3. The trimmed mean

$$T_1(\alpha) = \sum_{s=k+1}^{n-k-1} \frac{\overline{U}_s(1)}{n-2k-1},$$

where $\overline{U}_s(1)$ represent the ordered sample $U_s(1), s = \overline{1, n-1}$, $k = [(n-1)\alpha]$ – the whole part of number $(n-1)\alpha$, $\alpha \in [0, \frac{1}{2})$.

Notice that

3.1. $\alpha = 0,001$; 3.2. $\alpha = 0,05$; 3.3. $\alpha = 0,1$; 3.4. $\alpha = 0,25$.

Theorem. *The statistics 1-3 are unbiased and consistent in sense of a convergence in probability estimators for variogram* $2\gamma(1)$.

Proof. The proof of this theorem follows directly from the properties of expectation.

$$M\bar{U}(1) = \frac{1}{n-1}\sum_{s=1}^{n-1} MU_s(1) = 2\gamma(1).$$

Analogous calculation can be made for the median and trimmed mean.

$$M\tilde{U}(1) = 2\gamma(1);\ MT_1(\alpha) = 2\gamma(1),\ \alpha \in [0, \frac{1}{2}).$$

The consistent in sense of a convergence in probability estimators 1-3 for variogram is proved in [3].

Definition. *An estimator T_{n-1} is M-estimator, if*

$$inf_t \sum_{s=1}^{n-1} \eta(U_s(1) - t) = \sum_{s=1}^{n-1} \eta(U_s(1) - T_{n-1}),$$

where η is a real function [3].

Let $\eta'(x) = \psi(x)$. Then the estimator T_{n-1} is a solution of equation

$$\sum_{s=1}^{n-1} \psi(U_s(1) - T_{n-1}) = 0.$$

4. M-estimator T_n.

4.1. The Huber M-estimator, where

$$\psi(x) = \psi_1(x) = \begin{cases} x, & |x| \leq k, \\ \text{sgn}(x), & |x| > k, \end{cases}$$

$k = 1,4088$;

4.2. The Tukey M-estimator, where

$$\psi(x) = \psi_2(x) = \begin{cases} x(r^2 - x^2)^2, & |x| \leq r, \\ 0, & |x| > r, \end{cases}$$

$k = 4$;

4.3. The Hampel M-estimator, where

$$\psi(x) = \psi_3(x) = \begin{cases} x, & 0 \leq |x| \leq a, \\ a\text{sgn}(x), & a < |x| \leq b, \\ a\frac{r-|x|}{r-b}\text{sgn}(x), & b < |x| \leq r, \\ 0, & |x| > 14, \end{cases}$$

$a = 1,31, b = 2,039, r = 14$;

4.4. The Andrews M-estimator, where

$$\psi(x) = \psi_4(x) = \begin{cases} \sin x, & |x| \leq a\pi, \\ 0, & |x| > a\pi, \end{cases}$$

$a = 1,142$;

3 Distributions.

In order to illustrate the properties of variogram estimators 1-4, consider following distributions.

A. Standard normal N(0, 1);

B. N(0, 1) 5 per cent contaminated with N(0, 9):

$$0,95N(0,1) + 0,05N(0,9);$$

C. N(0, 1) 10 per cent contaminated with N(0, 9):

$$0,9N(0,1) + 0,1N(0,9);$$

D. N(0, 1) 25 per cent contaminated with N(0, 9):

$$0,75N(0,1) + 0,25N(0,9);$$

E. N(0, 1) 5 per cent contaminated with N(0, 100):

$$0,95N(0,1) + 0,05N(0,100);$$

F. Standard Laplace distribution with a density

$$f(x) = 0,5e^{-|x|};$$

G. Distribution with the density

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

Letting $\Phi(u)$ denote the cumulative distribution function of the standard normal distribution N (0, 1), the cumulative distribution function of a contaminated normal is defined by

$$F(u) = \epsilon\Phi(u) + (1-\epsilon)\Phi(u/3),$$

where $\epsilon = 0,95, \epsilon = 0,9$ and $\epsilon = 0,75$ in the case B-D, respectively, and

$$F(u) = \epsilon\Phi(u) + (1-\epsilon)\Phi(u/10),$$

$\epsilon = 0,95$ in the case E.

Using the procedure discussed in Huber [3], the variances of variogram estimators 1-4 for different distributions A - G may be computed. A program was developed in such computer algebra system as Mathematica. The values obtained from the program were listed in Table 1.

Table 1: The variances of 10 variogram estimators for 7 different distributions

	A	B	C	D	E	F	G
1	1,0000	1,4000	1,8000	3,0000	5,9500	2,0000	∞
2	1,5708	1,6810	1,8032	2,2620	1,7223	1,0000	2,4674
3.1	1,0003	1,3739	1,7749	2,9791	5,6217	1,9791	304,69
3.2	1,0263	1,1554	1,3173	2,1378	1,2462	1,6537	6,8470
3.3	1,0604	1,1688	1,2964	1,8490	1,2276	1,4941	3,8658
3.4	1,1952	1,2892	1,3949	1,8039	1,3285	1,2274	2,2732
4.1	1,0657	1,1649	1,2968	1,7877	1,2273	1,4406	2,9033
4.2	1,0989	1,1978	1,3107	1,7645	1,1780	1,4077	2,2593
4.3	1,0966	1,1954	1,3080	1,7603	1,1757	1,4558	2,3000
4.4	1,0998	1,1991	1,3125	1,7687	1,1789	1,4133	2,2687

4 Results.

The variances of 10 estimates of $2\gamma(1)$ for 7 different distributions are given in Table 1. We see that the mean is by far the most stable for the normal distribution, and by far the least stable in almost all of the heavy-tailed runs. All of the trimmed means were dominated by M-estimators only in case A. The estimators having consistently the smallest variances in the nonnormal cases are the M-estimators. Except for the contaminated case D. The median performs poorly.

Thus the M-estimators are the estimators of choice, having quite good efficiency for normal data coupled with stability for all the heavy-tailed distributions studied. The trimmed means and the median do not perform well.

References

[1] *Cressie N.* Statistics for Spatial Data. Wiley, New York (1991)

[2] *Matheron G.* Principles of Geostatisticss. J. Principles of Geostatistics **58** (1963) 1246-1266

[3] *Huber P.* Robustness in Statistics. Mir, Moscow (1984)(in Russian)