Robust estimation of the variogram in computer experiments

O. Roustant[1] , D. Dupuy[1], C. Helbert[1]

August 2007

[1] Ecole des Mines, Département 3MI, 158 Cours Fauriel, 42023 Saint-Etienne, France

Keywords: Computer experiments Variogram, Kriging model, Anisotropy, Robustness.

Extended abstract

In many industrial fields as oil engineering, aeronautics, automotive, nuclear engineering, complex phenomena are studied with simulators. These simulators are very useful to address various issues linked to the physical processes, from optimization to uncertainty evaluation. The drawback is that a simulation can last a long time: from 1 hour to several days (crash test, flow simulators). For this reason, there is a need to build proxy models for the simulators, on which the aforementioned issues can be treated. The corresponding methodology is known as DACE, for Design and Analysis of Computer Experiments, and was introduced in the 80's by (Sacks *et al.*, 1989 a and b). A computer experiment is another word for simulation.

The nature of computer experiments is different from physical experiments. For instance, if one neglects the numerical errors induced by convergence of algorithms or discretization variables for instance, simulations may be considered as deterministic[1]. In other words, the same input values will give the same output value. As a consequence, a model for the simulator should be obtained by interpolation techniques. The most famous one is known as kriging model. Initially introduced by Krige for mining engineering, the kriging model is very popular in computer experiments. The simulator output is considered as one realization of a Gaussian stochastic process ($Y(x)$):

$$Y(x) = m(x) + Z(x)$$

where x is a d-dimensional vector representing the inputs, $m(x)$ is a deterministic trend, and ($Z(x)$) a stationary centered stochastic Gaussian process with spatial correlation function $R(h)$.

A lot of information concerning $Y(x)$ is contained in the correlation function. In the following illustration, we represent several realizations of 1-dimensional kriging models, obtained with three different correlation functions. After trend estimation, the simulator may look like one of these realizations. We observe that the degree of smoothness is much different in the three cases.

[1] Except, of course, for a stochastic simulator, as in nuclear engineering.

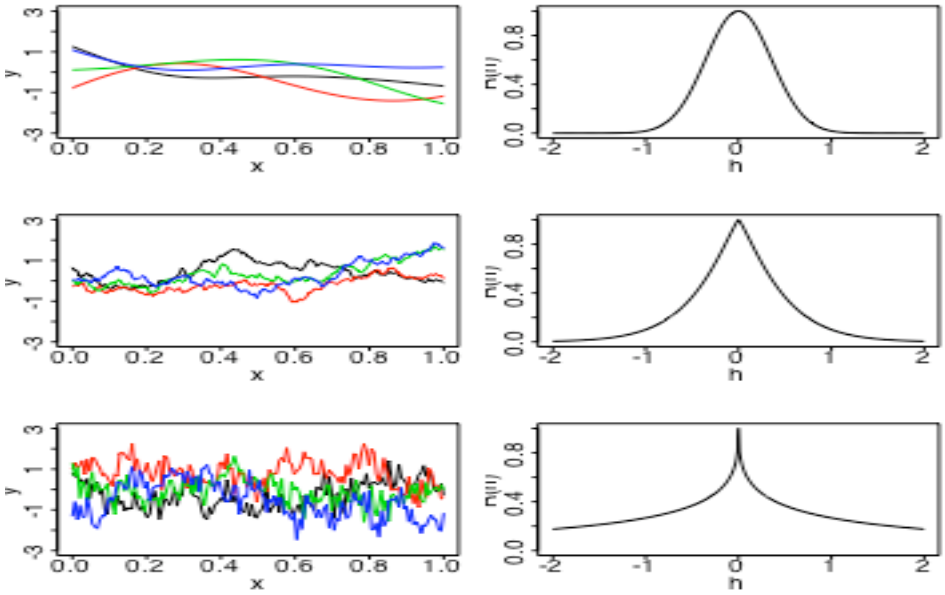

Figure 1. Three correlation functions (right) and 4 corresponding realizations (left). The correlation functions are $R(h) = \exp\left(-\left|\frac{h}{0.5}\right|^p\right)$ with *p*=2 (up), *p*=1.2 (middle) and *p*=0.4 (down).

Actually, the realizations were obtained with the same parametric form of correlation function, namely the power-exponential function:

$$R(h) = \exp\left(-\sum_{k=1}^{d} \theta_k |h_k|^{p_k}\right), \quad \text{with } 0 < p_k \le 2,\ k = 1,\ldots,d$$

which can be rewritten as $R(h) = \exp\left(-\sum_{k=1}^{d} \left|\frac{h_k}{\theta'_k}\right|^{p_k}\right)$, where the θ'_k are now scale parameters. The example corresponds to the one-dimensional case d=1. Knowing that there exist plenty of parametric forms for correlation functions, one can see that a kriging model can represent a lot of situations.

Since the shape of the simulator output is closely linked to the spatial correlation, a special care should be taken to the choice of the correlation function. In geostatistics, this a achieved by estimating the variogram

$$2\gamma(h) = \mathrm{var}[Z(x+h) - Z(x)]$$

Defined for intrinsic processes, the variogram is equivalent to $R(h)$ for stationary processes (see e.g. Cressie, 1993), and in this case we have $\gamma(h) = (1 - R(h)) \times \mathrm{var}[Z(x)]$ (this explains why a factor 2 is introduced in the above definition). Note that using the variogram instead of the correlation function is recommended even if the process is stationary, because of possible contaminations by trend estimate residuals (Cressie, 1993). Surprisingly, very little attention is paid to the selection of the correlation function in computer experiments: most often, a parametric form is pre-specified with no justification. Of course, the problem is much harder in computer experiments since the number of input variables is larger than the 2 or 3 considered in geostatistics. In addition, in our experience, an isotropic assumption is not realistic. But this is not a good reason to forgive all research on this important topic.

As a first step in this direction, we propose to compare three standard variogram estimators on a 1-dimensional example, suggested in the geostatistical litterature.

For sake of simplicity, we will assume no trend i.e. $Y(x) = Z(x)$ (in practise, the trend may be estimated and removed). We note $x^{(1)},...,x^{(n)}$ the *d*-dimensional data points (*n* different inputs for the simulator), and $z^{(1)},...,z^{(n)}$ the corresponding output values (obtained with the simulator). In the next example, we will have *d*=1.

The first variogram estimator is the method-of-moments estimator, proposed by Matheron (Cressie, 1993):

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}(z^{(i)} - z^{(j)})^2$$

where $N(h) = \left\{(x^{(i)},x^{(j)}),\ x^{(i)} - x^{(j)} = h\right\}$. In practise, the set *N(h)* is often empty and the above estimator have to be replaced by a local estimator of the mean. Matheron's estimator is not fully adapted to the problem. Indeed, the estimation of the variogram is a difficult statistical problem since the random variables $\left(Z(x+h) - Z(x)\right)^2$ are not independent and strongly skewed. In particular, large values may affect the estimation. For this reason, robust estimation is encouraged. For instance, one should replace the mean by a median in the definition above, or consider a a%-trimmed mean obtained by removing the a% highest and a% smallest values before computing the mean.

In geostatistics, two other estimators were proposed by Cressie-Hawkins (1980) and Genton (1998).

Cressie-Hawkins: $$2\hat{\gamma}_{Cressie-Hawkins}(h) = \frac{\left(\frac{1}{|N(h)|}\sum_{N(h)}\left|z^{(i)} - z^{(j)}\right|^{1/2}\right)^4}{0.457 + 0.494/|N(h)|}$$

Genton: $$2\hat{\gamma}_{Genton}(h) = Q_{N_h}{}^2,$$

where Q_{Nh} is 2.2191 times the *k*th quantile of $\{|z^{(i)} - z^{(j)}|,\ x^{(i)} - x^{(j)} = h\}$, with $k = \binom{\left[|N(h)|/2\right]+1}{2}/|N(h)|$, and $\left[|N(h)|/2\right]$ the integer part of $|N(h)|/2$.

The idea of Cressie and Hawkins was to transform the data in order to obtain a symmetrical distribution, which can be done by observing that the r.v. $\left|Z(x+h) - Z(x)\right|^{1/2}$ are nearly normally distributed. Note that their estimator is not robust to outliers, since it tends to infinity when only one $z^{(i)}$ tends to infinity. However, the term « robust » is usually attached to it. Actually, Cressie-Hawkins also proposed to replace the mean by the median, leading to a robust version of their estimator (Cressie-Hawkins, 1993, p. 75):

Cressie-Hawkins, robust version: $$2\tilde{\gamma}_{Cressie-Hawkins}(h) = \frac{med\left(\left|z^{(i)} - z^{(j)}\right|^{1/2}\right)^4}{0.457}$$

In next example, the results seem to be nearly the same with both versions. For this reason, we will work with the first definition.

The idea of Genton was to apply the results of scale estimators to the variogram estimation. Their definition can be found in (Rousseuw, Croux, 1993).

Finally, the constants 0.457, 0.494, 2.2191 are based on the normality assumption. These estimators may be sensitive to departures from this assumption.

We now compare the properties of these estimators on the 1-dimensional example proposed in (Genton, 1998). More precisely, we will consider 4 estimators: Matheron, Cressie-Hawkins, Genton and a 10% trimmed-mean (see last paragraph). The n data are obtained by simulation from a stationary centered Gaussian process with correlation function $R(h) = 1 - \gamma(h)$ corresponding to a spherical variogram:

$$\gamma(h) = a + b\left(1.5\left(\frac{h}{c}\right) - 0.5\left(\frac{h}{c}\right)^3\right)$$

with a=1, b=2, c=15. Then, ε% of the data (chosen at random) are removed and replaced by outliers, generated at random from a normal distribution $N(0,\sigma^2)$. Below, we give the results obtained with ε = 10%, 20% and $\sigma = 5$.

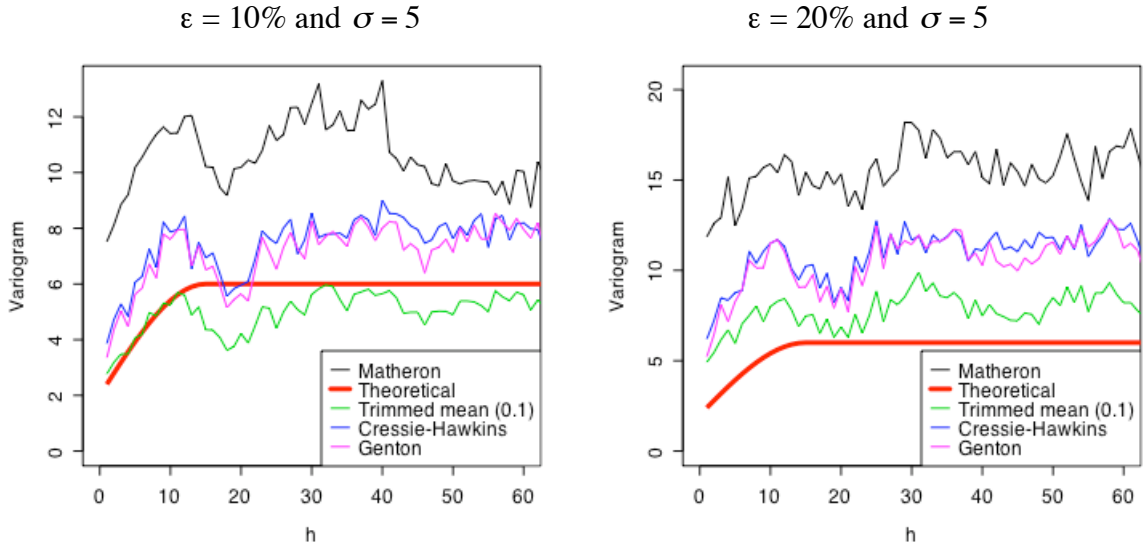

Figure 2. Comparison of standard variogram estimators in the presence ot outliers

We observe that all robust estimators behave better than Matheron's. The Cressie-Hawkins' estimator seems to behave as well as Genton one in this example, but both are outperformed by the trimmed mean. Finally, we see that in any case, it is not easy to guess a parametric form to the variogram. At least, with any of the three robust estimators, it is possible to retrieve the shape of the theoretical variogram. These conclusions differ significantly from Genton's.

As a conclusion to this brief study, we recommend to be very careful when using the robust estimators proposed in the geostatistical literature. First, on our example, they did not better than a trimmed-mean. In addition, we have tried to apply the robust estimators on real 3-dimensional data. In this case study, the trend was hard to capture. As a consequence, departures from normality were observed, due to trend contamination. The results of variogram estimation, obtained with Cressie-Hawkins' or Genton's estimators were impossible to exploit. Actually, these estimators seem to be very sensitive to departures from normality.

Finally, we see that the methodology of variogram estimation developed in geostatistics cannot be easily applied to computer experiments. It is even not clear if the variogram estimation is feasible in computer experiments, since the dimension if usually much larger than 2 or 3, and that the existing simplifying assumptions (isotropy for instance) are not realistic. However, this information is useful since the shape of the modeled output directly depends on it.

ACKNOWLEDGMENTS

This work was conducted within the frame of the DICE (Deep Inside Computer Experiments, http://dice-consortium.fr/) Consortium between ARMINES, Renault, EDF, IRSN, ONERA and TOTAL S.A.

References

Chilès J-P., Delfiner P. (1999), *Geostatistics. Modeling Spatial Uncertainty*, Wiley & Sons

Cressie N. (1993), *Statistics for Spatial Data*, Wiley & Sons

Cressie N., Hawkins D.H. (1980), "Robust estimation of the variogram: I", *Mathematical Geology*, 12 (2), 115-125

Genton M. (1998), "Highly Robust Variogram Estimation", *Mathematical Geology*, 30 (2), 213-221

Huber P.J. (1977), *Robust Statistical Procedures*, SIAM

Rousseuw P.J., Croux C. (1993), "Alternatives to the Median Absolute Deviation", *JASA*, 88 (424), 1273-1283

Sacks J., Schiller S.B., Welch W.J. (1989a), Designs for computer experiments, *Technometrics*, 31, pp. 41-47.

Sacks J., Welch W.J., Mitchell T.J., Wynn H.P. (1989 b), Design and Analysis of Computer Experiments, *Statistical Science*, 4, n°4, pp. 409-435.

Santner T.J., Williams B.J., Notz W.I. (2003). *The Design and Analysis of Computer Experiments*, Springer.