A SHORT NOTE ON VARIOGRAMS AND CORRELOGRAMS

Erik Mønness[1], Shirley Coleman[2]

1) erik.monness@hihm.no
   Hedmark University College,
   2450 Rena, Norway
   phone +47 62430598 / +47 62630510 fax +47 62430500
2) Shirley.Coleman@ncl.ac.uk
   ISRU, School of Mathematics and Statistics, Newcastle University,
   Newcastle upon Tyne, NE1 7RU, UK
   phone: +44 191 222 7096 fax +44 191 222 8020

Content

Context

The econometrics community advocates the use of the correlogram (Box, Jenkins et al. 1994). The geological community advocates the variogram (Webster and Oliver 2007). Both communities are interested in exploring auto dependence across series data. The correlogram is for a one-dimensional space, while the variogram is applicable in higher dimensions. But are they different on one-dimensional data? This short note explores the relationship between correlograms and variograms, illustrates their decomposition and considers the effect of approximations in their common use. Correlograms and variograms are an example of how obfuscation can occur when similar concepts are adopted by different specialists.

Consider a one-dimensional equally spaced series (e.g. in geology, a transect)

$X_i, \ldots, X_N$

Let $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\mathrm{var}_p(X) = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$

(NB: var_p for population variance, divided by N, not N-1)

Let

$$_{(N-k)}\overline{X} = \frac{1}{N-k}\sum_{i=1}^{N-k} X_i \quad \text{and} \quad \overline{X}_{(N-k)} = \frac{1}{N-k}\sum_{i=1}^{N-k} X_{i+k}$$

These are the means of the first and last N-k values in the series, respectively. Equally, let

$$\mathrm{var}_p({}_{(N-k)}X) = \frac{1}{N-k}\sum_{i=1}^{N-k}\left(X_i - {}_{(N-k)}\overline{X}\right)^2 \quad \text{and} \quad \mathrm{var}_p(X_{(N-k)}) = \frac{1}{N-k}\sum_{i=1}^{N-k}\left(X_{i+k} - \overline{X}_{(N-k)}\right)^2$$

An example time series is shown in Figure 1. The large dip is the financial crisis in 2008.

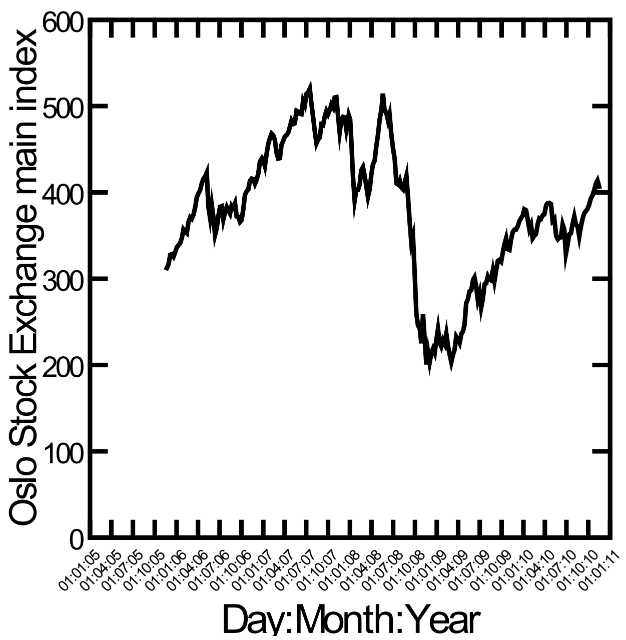

Figure 1. Oslo Stock Exchange main mean weekly Index from November 2005 to November 2010

The Correlogram

The usual way of presenting a correlogram 1 to K is to assume N is large and K<<N so that the autocovariance is approximately given by c(k) where

$$c(k)=\frac{1}{N-k}\sum_{i=1}^{N-k}\left(X_{i+k}-\bar{X}\right)\left(X_i-\bar{X}\right) \quad \text{for } k=0,\ldots,K$$

Dividing the autocovariance by $var_p(X)$ which is c(0), gives the k^{th} autocorrelation

$$r(k)=\frac{c(k)}{c(0)}$$

A plot of r(k) against k is referred to as a correlogram. Under a model with stable mean and variance, «stationary», one can test if the autocorrelation is zero or not. The assumption of stationarity is often based on an inspection of the data. However, the statistic can be calculated without any assumptions.

The exact autocovariance is $c'(k)=\frac{1}{N-k}\sum_{i=1}^{N-k}\left(X_{i+k}-\bar{X}_{(N-k)}\right)\left(X_i-{}_{(N-k)}\bar{X}\right)$

This is just an ordinary covariance between two sets of data, which may or may not be overlapping.

The Variogram

Like the correlogram, variograms also look at variation between successive points but are less widely used, and were originally proposed in the context of soil science and gold mining. The «semi variance» in a one-dimensional series is defined as

$$\gamma(k)=\frac{1}{2}\frac{1}{N-k}\sum_{i=1}^{N-k}\left(X_{i+k}-X_i\right)^2$$

A plot of γ(k) (or γ(k)/c(0), as used here) against k is called a variogram. Often it is assumed that $\left(X_{i+k}-X_i\right)$ is a stationary process. This is a weaker condition on the X's than the one in use with autocorrelation. γ(k) can also be seen as half the unexplained variance after a regression of $X_{i+k}=0+1*X_i$.

Same or equal?

Adding the exact autocovariance and semi variance yields

$$c'(k)+\gamma(k)=\frac{1}{2}\left\{\mathrm{var}_p\left(X_{(N-k)}\right)+\mathrm{var}_p\left({}_{(N-k)}X\right)+\left(\overline{X}_{(N-k)}-{}_{(N-k)}\overline{X}\right)^2\right\}$$

Assuming the approximation for autocovariances, if the lag is small then $c'(k)+\gamma(k)\approx c(0)$ and $c'(k)\approx c(k)$ yielding $r(k)=\frac{c(k)}{c(0)}\approx 1-\frac{\gamma(k)}{c(0)}$

The correlogram is therefore approximately equal to 1 – variogram and both are considered to give more or less the same information.

However, when the lag is large, the two parts of the series become detached, and using the exact autovariance, $c'(k)+\gamma(k)$ can have any positive value including c(0), and the correlogram and variogram no longer sum to unity.

The Oslo Stock Exchange data are not stationary but a first difference of their values is stationary. In Figure 2 exact calculations of correlogram and variogram are performed on the first difference of the Oslo Stock Exchange main mean weekly Index from November 2005 to November 2010. The correlogram is usually calculated for K<<N. As N is 260 for the Oslo Stock Exchange data, Figure 2 compares the correlogram and variogram for the first 30 lags (out of 260).

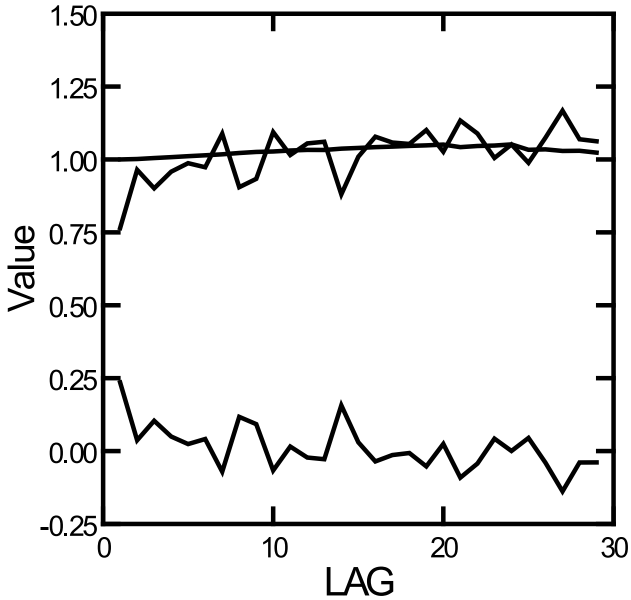

Figure 2. Correlogram (bottom), variogram (top) and their sum (smooth line)

The «sill» appears to occur at lag 2. The sum of correlogram and variogram for each lag is approximately unity. In Figure 3, the graph is extended to all lags.

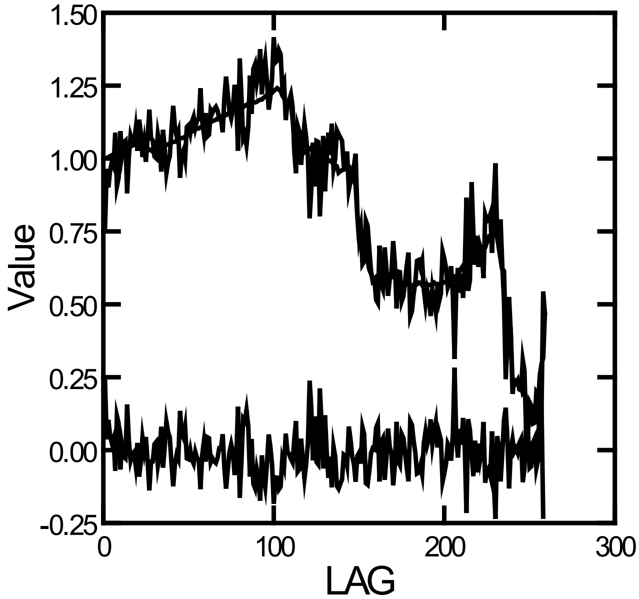

Figure 3. Correlogram (bottom), variogram (top) and their sum (smooth line hidden within the variogram)

The variogram starts out at zero and grows to its maximum at about the 100th week. The exact autocorrelation starts out at 1, decreasing to its minimum negative value at week 100. Both the variogram and the correlogram alternate in an opposite fashion. The sum has a fairly smooth curve but already deviates from unity at moderate lags. The sum grows until about lag 100 and then decreases ending up at approximately zero.

Decomposition

Semi variances can be seen as a decomposition of the total mean square error (MSE) (Bachmaier and Backes 2008):

$$\mathrm{MSE} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 = \frac{1}{2}\frac{1}{N(N-1)}\sum_{i \neq j}\left(X_i - X_j\right)^2 = \frac{1}{N(N-1)}\sum_{i<j}\left(X_i - X_j\right)^2$$

$$= \frac{1}{N(N-1)}\sum_{k=1}^{N-1}\sum_{i=1}^{N-k}\left(X_{i+k} - X_i\right)^2 = \frac{2}{N(N-1)}\sum_{k=1}^{N-1}(N-k)\gamma(k).$$

The MSE is the mean of every squared difference $\left(X_i - X_j\right)^2$. It can be observed that the semi variance, γ(k), is the mean of the squared differences on the kth diagonal in the $\square^2$ space of squared differences. The MSE is approximately equal to c(0), and so each autocovariance is approximately equal to the mean of the other differences. The decomposition is illustrated in the contour map in Figure 4 using absolute differences rather than squared differences as their size distorts the map.

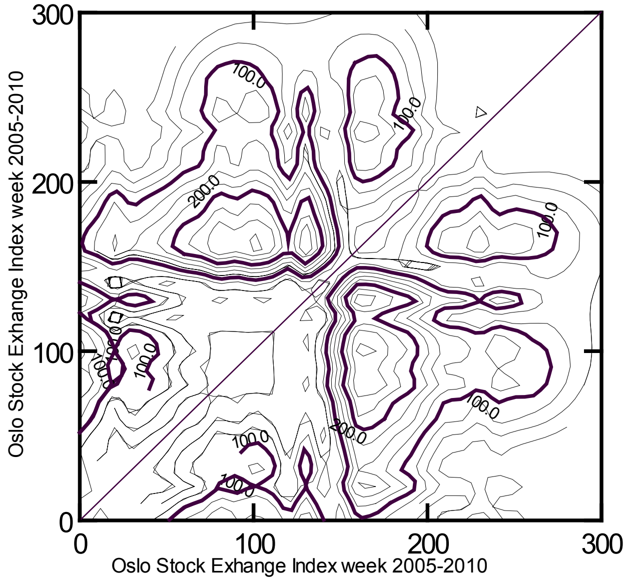

Figure 4. Contour map of absolute differences in the Oslo Stock Exchange main weekly Index from November 2005 to November 2010

The contour map depicts the landscape of absolute differences between weekly index values. The main diagonal has height zero. From the main diagonal, one can start a walk horizontally (or vertically) to experience the change from that week. It can be seen that around week 170, the differences are very large corresponding to the financial crisis. The semi variances are the mean of squared values along diagonals parallel to the main diagonal.
The variogram can be applied relaxing the need of equally spaced data, and can be extended to higher dimensions.

Conclusion.

The variogram and correlogram are closely related measures applied to a one-dimensional series. They are examples of different communities developing different measures for nearly the same thing. They do, however, differ for larger lags. Both semi variances and autocovariances can be obtained from the decomposition of the mean square error. The components of the decomposition can be illustrated in a contour plot which highlights areas of large variation in the series data. The variogram can be applied relaxing the need of equally spaced data, and can be extended to higher dimensions.

References

Bachmaier, M. & Backes, M. (2008). Variogram or semivariogram? Understanding the variances in a variogram. *Precision Agriculture 9*(3), 173-175. doi: 10.1007/s11119-008-9056-2

Box, G. E. P., Jenkins, G. M. & Reinsel, G.C. (1994). *Time series analysis: Forecasting and control.* Englewood Cliffs, N.J.: Prentice-Hall.

Webster, R. & Oliver, M. A. (2007). *Geostatistics for environmental scientists.* Wiley & Son.