University of California, Los Angeles
Department of Statistics

Statistics C173/C273 **Instructor: Nicolas Christou**

The variogram

- Let $Z(s)$ and $Z(s+h)$ two random variables at locations s and $s+h$. Intrinsic stationarity is defined as follows:

$$E(Z(s+h) - Z(s)) = 0$$

and

$$Var(Z(s+h) - Z(s)) = 2\gamma(h)$$

The quantity $2\gamma(h)$ is known as the variogram and is very crucial in geostatistics. The variogram says that differences of variables lagged h-apart vary in a way that depends only on h through the length of h. This is called *isotropic* variogram as opposed to *anisotropic* variogram which depends not only on the length h but also the direction. Because of the assumption of constant mean (no trend) we have $E(Z(s)) = \mu$ and we can write

$$\text{Var}(Z(s+h) - Z(s)) = E(Z(s+h) - Z(s))^2$$

Therefore we can use the method of moments estimator for the variogram (also called the classical estimator):

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,$$

where the sum is over $N(h)$ such that $s_i - s_j = h$.

Robust estimator:
Cressie and Hawkins (1980) proposed the following estimator for the variogram which is robust to outliers compared to the classical estimator:

$$2\bar{\gamma}(h) = \frac{\left\{\frac{1}{N(h)} \sum_{N(h)} |Z(s_i) - Z(s_j)|^{\frac{1}{2}}\right\}^4}{0.457 + \frac{0.494}{N(h)}}$$

where the sum is over $N(h)$ such that $s_i - s_j = h$.

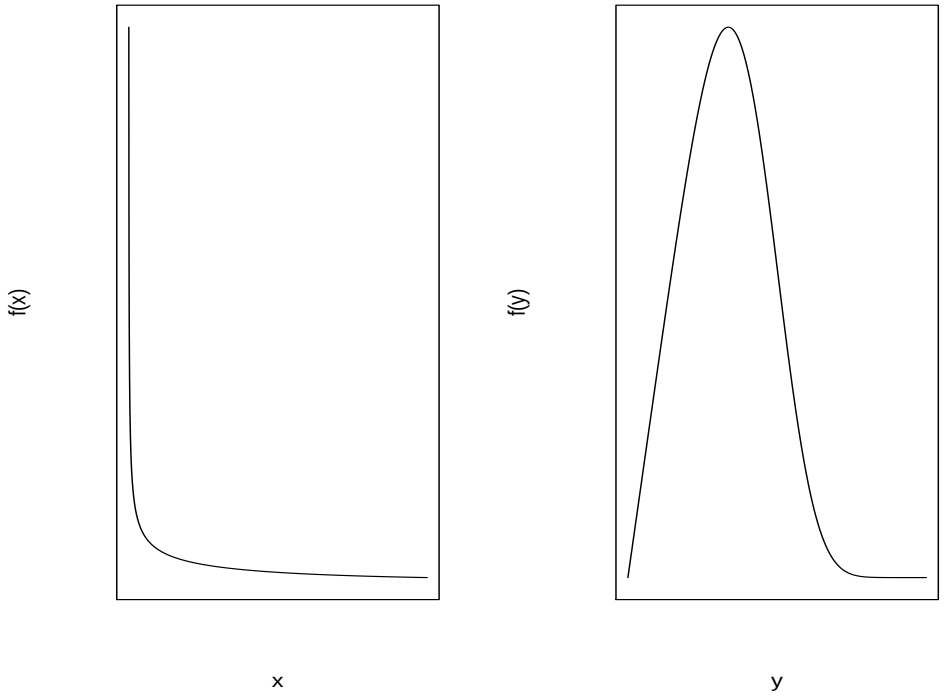
The idea behind the robust estimator is described below: If the process $Z(s)$ follows the normal distribution then

$$Z(s+h) - Z(s) \sim N\left(0, \sqrt{2\gamma(h)}\right)$$

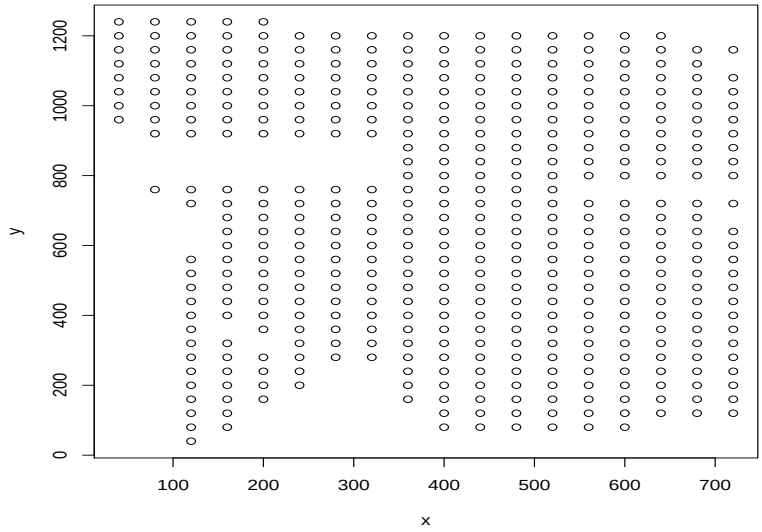
Therefore:

$$\left(\frac{Z(s+h) - Z(s)}{\sqrt{2\gamma(h)}}\right)^2 \sim \chi_1^2 \quad (\chi^2 \text{ distribution with 1 degree of freedom}).$$

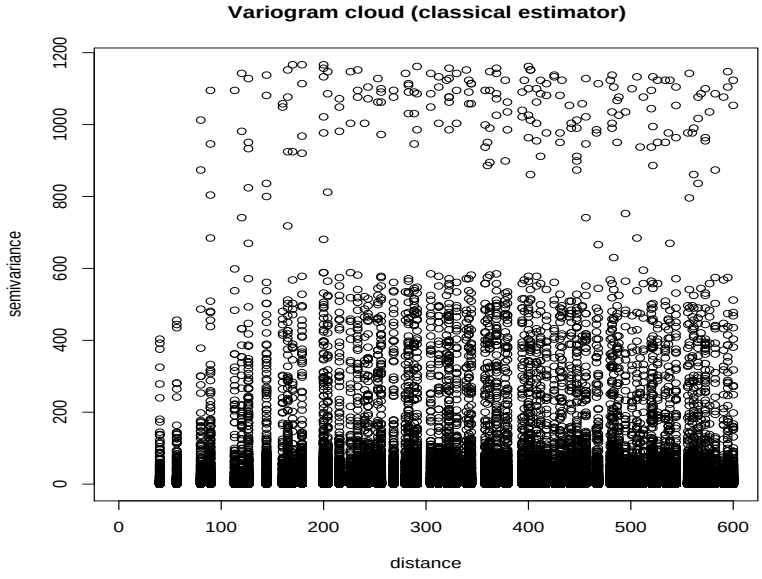
This is a highly skewed distribution. However if $X \sim \chi_1^2$ then $Y = X^{\frac{1}{4}}$ has an approximately symmetric distribution (see figure below). We would expect that the quantity $(Z(s+h) - Z(s))^{\frac{1}{2}}$ will behave much better than $(Z(s+h) - Z(s))^2$.



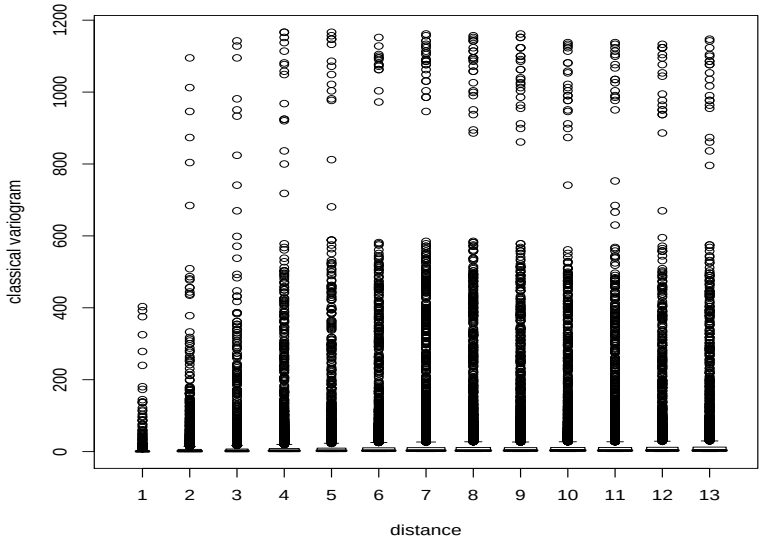
- **The variogram cloud:**
  For the example below we will compute the square of the difference between each pair separated by distance h and plot these values against the values of h. This is called a variogram cloud. The data concern top soil phosphorus in mg per liter for data collected at the Broom Barn Farm. The variables measured included pH, exchangeable potassium (K) in mg per liter, and available phosphorus (P), also in mg per liter. Here are the 434 data points (distance between points north-south and east-west is 40 m):



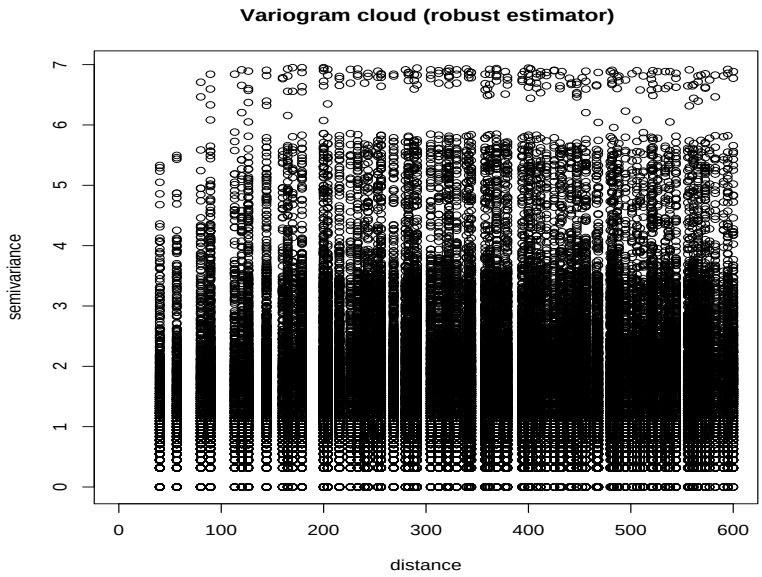
And here is the variogram cloud (classical estimator):



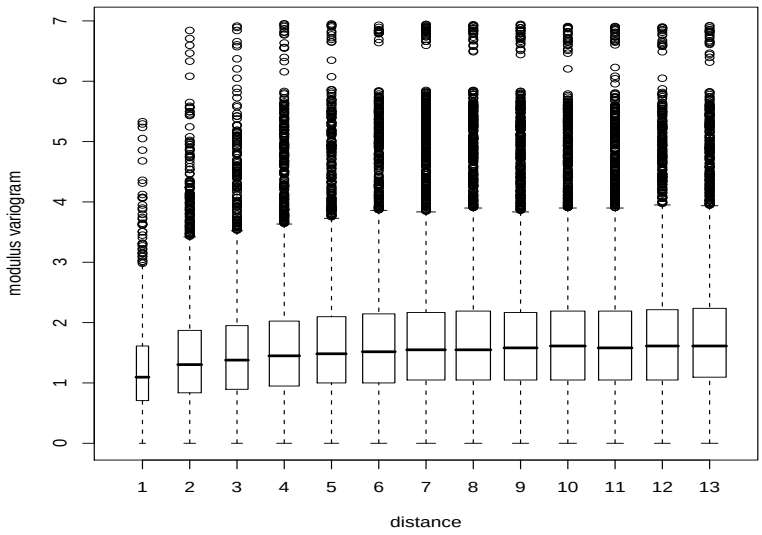
Box plots can be constructed using the variogram cloud of the classical estimator above:



Using the robust estimator the variogram cloud we obtained is the one below:



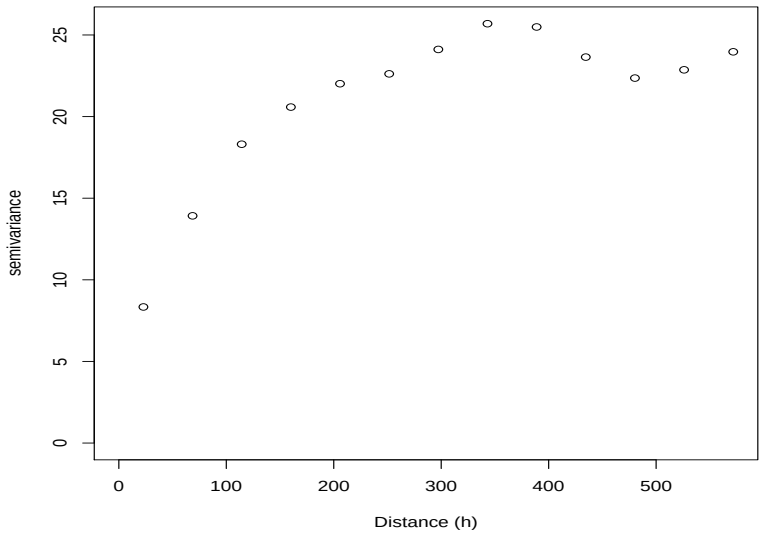
And here are the box plots using the variogram cloud of the robust estimator above:



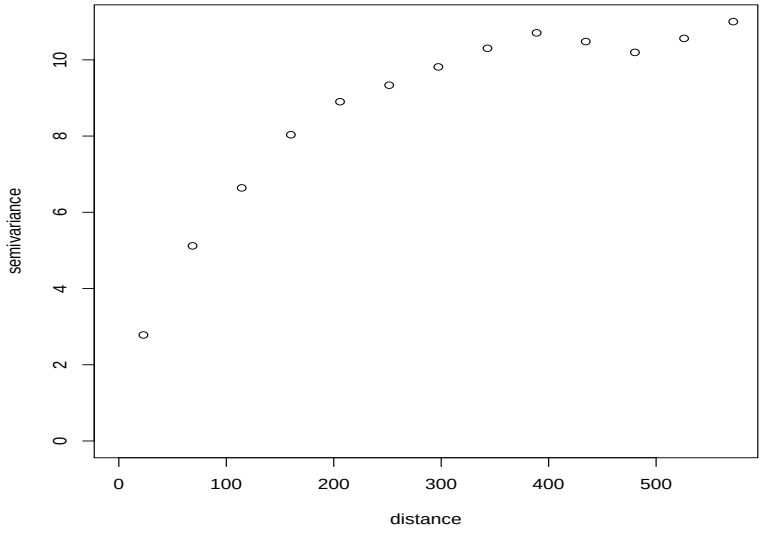
We observe that the robust estimator is much more robust to outliers. Of course in this example using logarithms can improve the picture even further.

However it is very difficult to judge from the variogram cloud if there is any spatial correlation. Instead, the average of the points for each separation distance h is computed to get the graphs below.

The semivariogram plot (classical estimator):



The semivariogram plot (robust estimator):



- **Modeling the sample variogram**
  Once the sample variogram is computed, a function is fit to it. In other words, we try to come up with what the variogram graph would look like if we had the entire population of all possible pairs. Popular variogram models that are used are the linear, spherical, and exponential (also see graphs on next pages).

  Linear model:
  This is the simplest model for the semivariogram graph. It depends on one parameter, the slope b.

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + bh, & h \neq 0 \end{cases}$$

$$\theta = (c_0, b)', \text{ where } c_0 \geq 0 \text{ and } b \geq 0.$$

  Spherical model:
  This is the model that proposed by Matheron. It has two parameters: The range of influence and the sill (or plateau) which the graph reaches at distances h larger then the range. Generally, the range of influence is the distance beyond which pairs are unrelated.

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + c_1(\frac{3}{2}(\frac{h}{\alpha}) - \frac{1}{2}(\frac{h}{\alpha})^3), & 0 < h \leq \alpha \\ c_0 + c_1, & h \geq \alpha \end{cases}$$

$$\theta = (c_0, c_1, \alpha)', \text{ where } c_0 \geq 0,\ c_1 \geq 0, \text{ and } \alpha \geq 0.$$

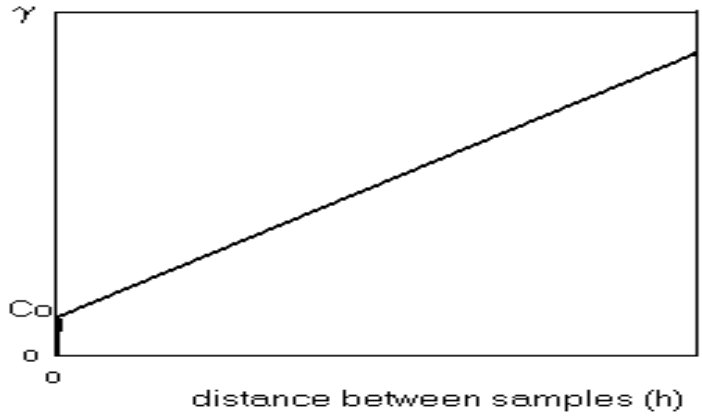
  Exponential model:
  This model represents an exponential decay of influence between two sample (larger distance between two samples means larger decay). It depends on two parameters, the range and the sill (plateau).

$$\gamma(h;\theta) = \begin{cases} 0, & h = 0 \\ c_0 + c_1(1 - exp(-\frac{h}{\alpha})), & h \neq 0 \end{cases}$$

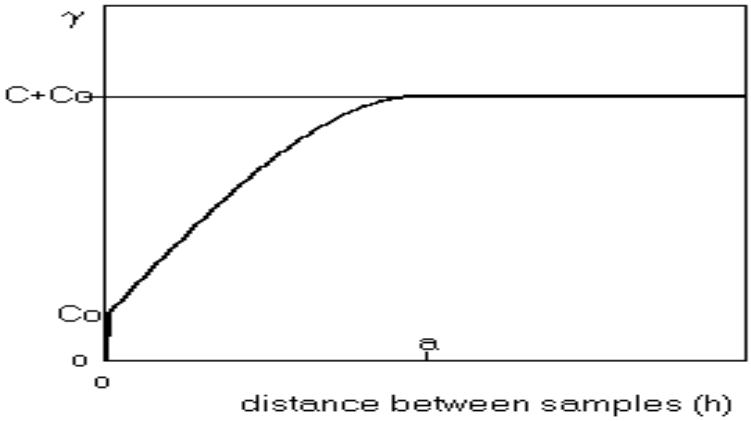
$$\theta = (c_0, c_1, \alpha)', \text{ where } c_0 \geq 0,\ c_1 \geq 0, \text{ and } \alpha \geq 0.$$

  Other models:
  There are other models, such as, the Gaussian model, the hole effect model, the Paddington mix model, the circular model, cubic model, Matérn function, etc.

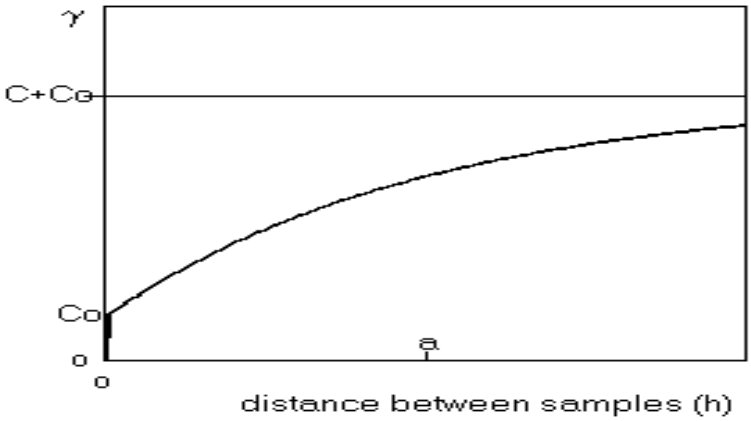
Linear semivariogram:



Spherical semivariogam:



Exponential semivariogam:



Estimation of variogram parameters:
This may not be easy! Different approaches were proposed during the last 40 years to fit a model to the sample variogram and estimate the variogram's parameters. As a compromise between efficiency and simplicity, Cressie (1985) advocates minimizing a weighted sum of squares

$$\sum_{k=1}^{K} \left\{ \frac{2\hat{\gamma}(\boldsymbol{h}(k))}{2\gamma(\boldsymbol{h}(k);\boldsymbol{\theta})} - 1 \right\}^2 |N(\boldsymbol{h}(k)|$$

with respect to variogram parameters θ. The sequence $\boldsymbol{h}(1), \boldsymbol{h}(2), \cdots, \boldsymbol{h}(k)$ denotes the lags at which the classical estimator was computed. Zimmermann and Zimmerman (1991) summarize and compare several methods of variogram-parameter estimation. They find that the weighted-least-squares approach usually performs well, and never does poorly, against other competitors.

Variogram model parameters
The various parameters of the variogram model are:

1. Nugget Efffect (c_0):
   If we stand by the assumption that sample values are measured precisely and accurately then the semi-variogram model must have a value of zero at zero distance. It is like calculating the difference of $Z(s)$ with itself. That is,

   $$\gamma(0) = 0$$

   The term nugget effect (or nugget variance) was introduced on the basis of the interpretation of gold mineralization. It was suggested by Matheron (1962) and it is believed that microscale variation (small nuggets) is causing a discontinuity at the origin.

2. Range (α):
   As the separation distance increases the value of the variogram increases as well. However, after a certain distance the variogram reaches a plateau. The distance at which the variogram reaches a plateau is the range.

   We generally interpret the range of influence as that distance beyond which pairs of sample values are unrrelated. Beyond the range the variogram remains essentially constant.

3. Sill ($c_0 + c_1$):
   It is the variogram value for separation distances $h \geq \alpha$.

Outline of spatial continuity analysis:

1. We begin usually with the calculation of an *omnidirectional* variogram. With all possible directions combine in a single variogram only the separation distance is important. The omnidirectional variogram can be thought as the average variogram of all directions. Strictly speaking is not the average because in one direction we may have more pairs than other directions.

2. The second step is to explore anisotropy by calculating the *directional* variograms for different directions (one at a time). In many spatial data the direction of anisotropy may be determined by the nature of the problem. For example, if we analyze airborne pollutants, the wind direction may be an important factor in the calculation of the variogram.

3. Once we decide which directional variogram we want to calculate, we must choose the distance parameters. There are two distance parameters. The first one is the separation distance (h). If the samples form approximately a grid, then the grid distance can be a good choice for h. If the samples do not form a regular grid the separation distance is chosen to be equal to the average distance of neighboring samples. The second distance parameter is the tolerance we allow on h. This will give us enough pairs for the variogram calculation. The common choice for the tolerance is half of the separation distance (h). For example, if we use $h = 10\ m$ then we will use all the points that fall between $5\ m$ and $15\ m$, etc.

4. Another parameter that we need to choose is the *angular* tolerance. When we calculate directional variograms ideally we want to use small angular tolerances so that the direction is preserved. However, many times we need to choose an angular tolerance large enough to produce enough number of pairs for the variogram calculation. We can try different angular tolerances and use the smallest one that produces good results.