STA 5377, Homework 1

Carson Slater *Baylor University*

1

(Answer this question without using the computer.) Consider the following (very small) data set:

Vector #	1	2	3	4	5
x	1	1	2	2	4
y	1	3	1	3	3
z	10	10	12	110	15

(a)

Compute and plot the sample semivariogram $\hat{\gamma}$. In doing so, use every possible distance h. That is, do not bin the h values.

The classic semivariogram estimator is:

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

where $N(h)$ is the humber of paris of locations that are h distance apart. We compute the distances between each vector:

	1	2	3	4
2	2			
3	1	2.236		
4	2.236	1	2	
5	3.606	3	2.828	2

So for $h = 1$, the sample variogram be

$$\hat{\gamma}(1) = \frac{1}{2(2)}(10 - 12)^2 + (10 - 110)^2 = 2501.$$

For $h = 2$, the sample variogram be

$$\hat{\gamma}(2) = \frac{1}{2(3)}(10 - 10)^2 + (12 - 110)^2 + (110 - 15)^2 = 3104.833.$$

For $h = 2.236$, the sample variogram be

$$\hat{\gamma}(2.236) = \frac{1}{2(2)}(110 - 10)^2 + (12 - 10)^2 = 2501,$$

For $h = 2.828$, the sample variogram be

$$\hat{\gamma}(2.828) = \frac{1}{2}(12 - 15)^2 = 4.5,$$

For $h = 3$, the sample variogram be

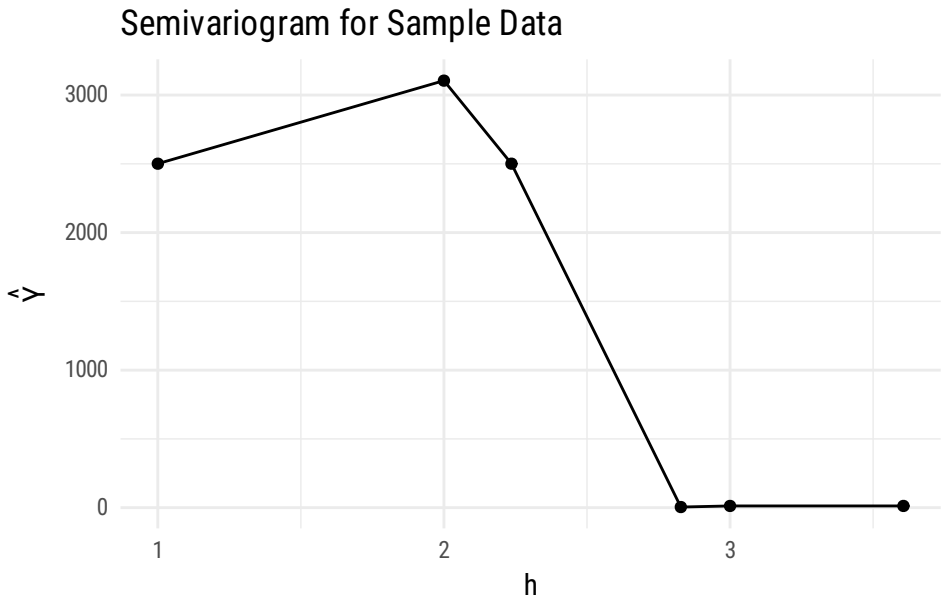
$$\hat{\gamma}(3) = \frac{1}{2}(10 - 15)^2 = 12.5,$$

For $h = 3.606$, the sample variogram be

$$\hat{\gamma}(3.606) = \frac{1}{2}(10 - 15)^2 = 12.5,$$

These calculations yield the following:

h	1	2	2.236	2.828	3	3.606
$\tilde{\gamma}(h)$	2501	3104.833	2501	4.5	12.5	12.5



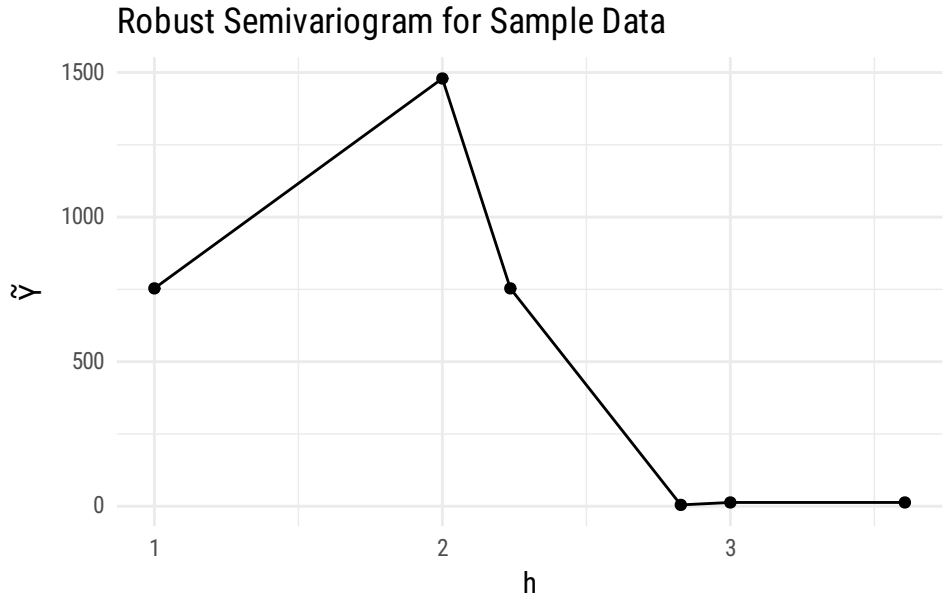
(b)

Compute and plot the robust sample semivariogram $\tilde{\gamma}$, again using every possible distance h. The robust semivariogram estimator is:

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

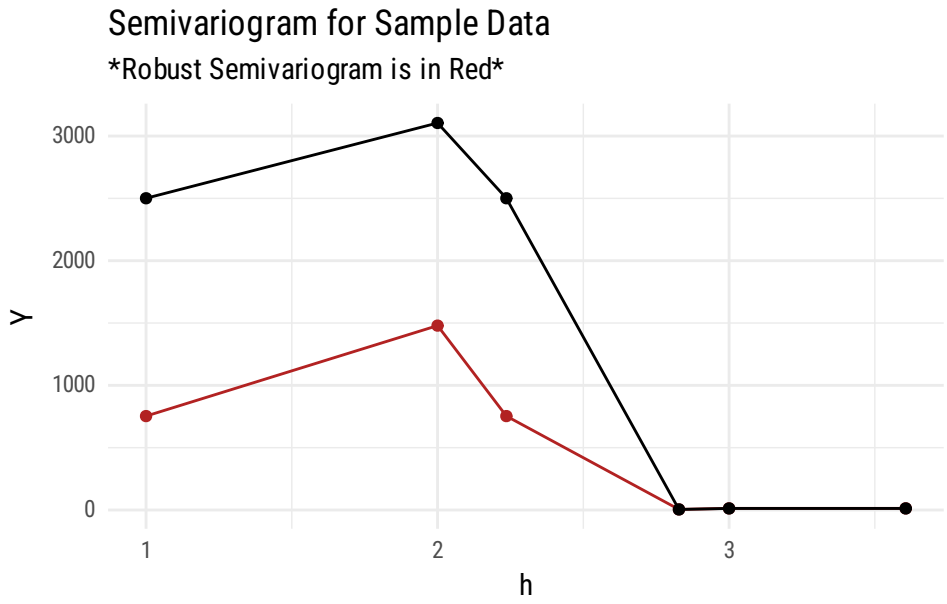
For the sake of brevity, calculations are not shown but we have the following robust sample semivariogram $\tilde{\gamma}$:

h	1	2	2.236	2.828	3	3.606
$\hat{\gamma}(h)$	753.462	1479.277	753.462	4.731861	13.14406	13.14406



(c)

Compare the behavior of the robust and standard estimates of γ.

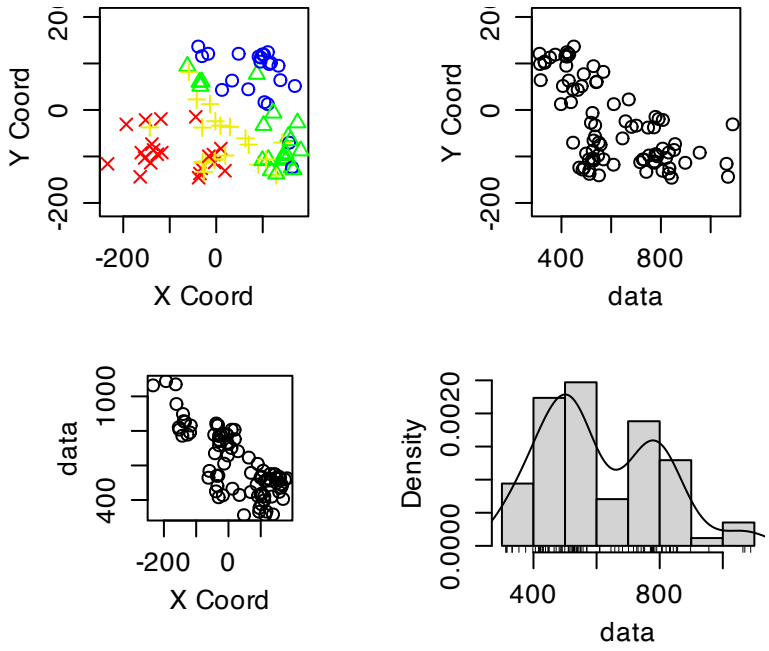


The robust estimator seems to dampen extreme values and elevate the smallers ones, but not by much it seems.

2

The `wolfcamp` data set is included in the package `geoR`. The data consist of the locations of 85 wells in the Wolfcamp aquifer in Texas. For each well, the piezometric head was measured; essentially, this is the level of the water table above sea level. `wolfcamp$coord` contains the locations of the observations; `wolfcamp$data` contains the Z values. Plot the data using `plot.geodata` and explain your findings.

```
library("geoR")
plot.geodata(wolfcamp)
```



For the spatial plot (top right), there appears to be clustering for binned measurments of z values. Both the x and y coordinates seem to be negatively correlated with the z values. Additionally, the z values appear to have a bimodal distribution.

3

For $\{Z(s) \ : \ s \in D\}$, what is the variance of $Y = \sum_{i=1}^{n} a_i Z(s_i)$, where $a_1, \ldots, a_n$ are real constants?

Using the properties of variance and covariance, we have:

$$\text{Var}(Y) = \text{Var}\left(\sum_{i=1}^{n} a_i Z(s_i) \right).$$

By expanding the variance of a linear combination,

$$\text{Var}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \text{Cov}(Z(s_i), Z(s_j)).$$

In matrix notation, let $\mathbf{a} = (a_1, \ldots, a_n)'$ be the vector of coefficients, and let $\mathbf{C}$ be the $n \times n$ covariance matrix with entries $\text{Cov}(Z(s_i), Z(s_j))$. Then,

$$\text{Var}(Y) = \mathbf{a}'\mathbf{C}\mathbf{a}.$$

4

The covariance function of $\{Z(s) \ : \ s \in D\}$ is

$$C(h) = \text{Cov}(Z(s), Z(s+h)) = \mathbb{E}\{[Z(s) - \mu(s)][Z(s+h) - \mu(s+h)]\}.$$

Explain why $C(h)$ must be a nonnegative function.

If $\{s_j\}$ is any collection of locations, then complex linear combinations

$$\mathbf{a}'(Z - \mu) = \sum a_j (Z_j - \mu_j)$$

of the centered random variables $Z_j = Z(s_j)$ (with means $\mu_j = \mu(s_j)$) must have nonnegative squared modulus

$$\mathbb{E}\left|\sum a_j (Z_j - \mu_j)\right|^2 = \sum a_j C(s_j - s_k)\overline{a_k} \geq 0$$

for every set of complex numbers $\{a_j\} \subset \mathbb{C}$. Recall that $\overline{a_k}$ is the complex conjugate of a_k. Then, a function $C_0(h)$ is called *positive semi-definite* if it always satisfies the inequality

$$\sum_{jk} a_j C(s_j - s_k)\overline{a_k} \geq 0 \tag{1}$$

for any locations s_j and complex numbers a_j. Note how (1) is the analogue of the result we found in question 3. This is equivalent to asking that

$$C(h) = C(-h)$$

for every $h \in \mathbb{R}^n$ and that

$$\sum a_i C(s_i - s_k) a_k \geq 0$$

for all *real* numbers $a_i \in \mathbb{R}$. This implies the covariance function is a positive semidefinite function, or alternatively a nonnegative function.