working with spatial data

Kriging (Part I) · also referred to as Gaussian Processes and Spatial Regression

Beginning with the three items obtained from the Variogram:

- $K^{\text{neighbors}}$ matrix containing the estimated covariance between pairs of training points
- K^{x} vector containing estimated covariance between new points and each training point
- an estimate of the covariance $k(x,x)$

$$\begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ k(x_n,x_1) & \cdots & \cdots & k(x_n,x_n) \end{bmatrix} \qquad \begin{bmatrix} k(x,x^1) \\ k(x,x^2) \\ \vdots \\ k(x,x^n) \end{bmatrix} \qquad k(x,x)$$

$$\boldsymbol{K^{\text{neighbors}}} \qquad \boldsymbol{K^{x}}$$

$$\begin{bmatrix} 1.9 & 1.2 & \cdots & 0.005 \\ 1.2 & 1.9 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0.005 & \cdots & \cdots & 1.9 \end{bmatrix} \qquad \begin{bmatrix} 0.5 \\ 0.6 \\ \vdots \\ 0.2 \end{bmatrix} \qquad k(x,x) = 1.9$$

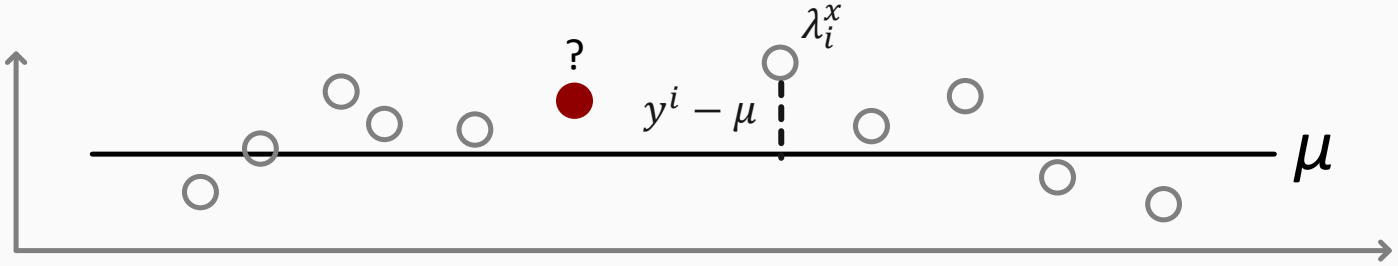
Two ways to derive this method lead to the same result; Maximum Likelihood (MLE) or Optimization

Maximum Likelihood Estimation is often thought as Kriging and Optimization as Gaussian Processes.

$$\hat{\gamma}(x) = \mu + \sum_{i \text{ neighbours}} \lambda_i^x \cdot (y^i - \mu)$$

The estimation is anything is varying in the sample space (label at point x). The estimation is the trend μ plus the summation over the residuals (distances between neighbours) $(y^i - \mu)$ weighted λ_i^x for the neighbours that depend on x. Highly correlated neighbors do not all have to be taken into account.

The question is: How to get a prediction for y at this point?



Examine the neighbors and determine how far above the trend each is $(y^i - \mu)$, assuming the distance is about the same as all the neighbours. The weighted average is computed of the distances from the trend λ_i^x.

The remaining questions: How to compute mu μ and the lambda λ for a particular x.

Kriging (Part II) · also referred to as Gaussian Processes and Spatial Regression

$$\hat{y}(x) = \mu + \sum_{i \text{ neighbours}} \lambda_i^x \cdot (y^i - \mu)$$

Mu μ is simple to compute in terms of Kriging as it estimates the trend by the empirical mean of all the examples. However, the mean μ can easily be removed by normalizing everything to 0 and assuming the mean is then 0.

Determining lambda λ in terms of Kriging is more involved:

Define prediction error as the predicted value less the actual value: $\hat{y}(x) - y(x)$, Kriging functions to minimize the prediction error.

Prediction Error intuition: $\hat{y}(x) - y(x) = (\hat{y}(x) - \mu) - (y(x) - \mu)$ to get distances from the mean. These distances in turn, can be referred to as remainders: $(\hat{R} - R)$ which is the remainder of the empirical (estimated) value and the actual value.

Knowing that Kriging tries to minimize the prediction error $(\hat{R} - R)$ so the variance of the prediction error $\hat{\sigma}^2(x)$ is as follows:

$$\hat{\sigma}^2(x) = \text{Var}(\hat{R}) + \text{Var}(R) - 2\text{Cov}(\hat{R}, R)$$

Plugging in the variables from above:

$$\hat{\sigma}^2(x) = \sum_i \sum_k \lambda_i^x \lambda_k^x (x^i, x^k) + k(x,x) - 2\sum_i \lambda_i^x k(x^i, x)$$

The method Kriging minimizes variance is by setting the derivatives to 0:

$\frac{\partial \hat{\sigma}^2(x)}{\partial \lambda_i^x} = 2\sum_k \lambda_k^x k(x^1, x^k) - 2k(x^i, x) = 0$ → Taking the partial derivative of the variance $\hat{\sigma}^2$ with respect to the first component of lambda λ.

The expression can be vectorized as follows for setting all derivatives to 0:

$\lambda^x K^{\text{neighbors}} = K^x$ and solve for λ with an inverse: $\lambda^x = K^x K^{\text{neighbors}^{-1}}$

$\mathbb{R}^{1 \times \text{neighbours}}$ $\quad$ $\mathbb{R}^{\text{neighbours} \times \text{neighbours}}$ $\quad$ $\mathbb{R}^{1 \times \text{neighbours}}$

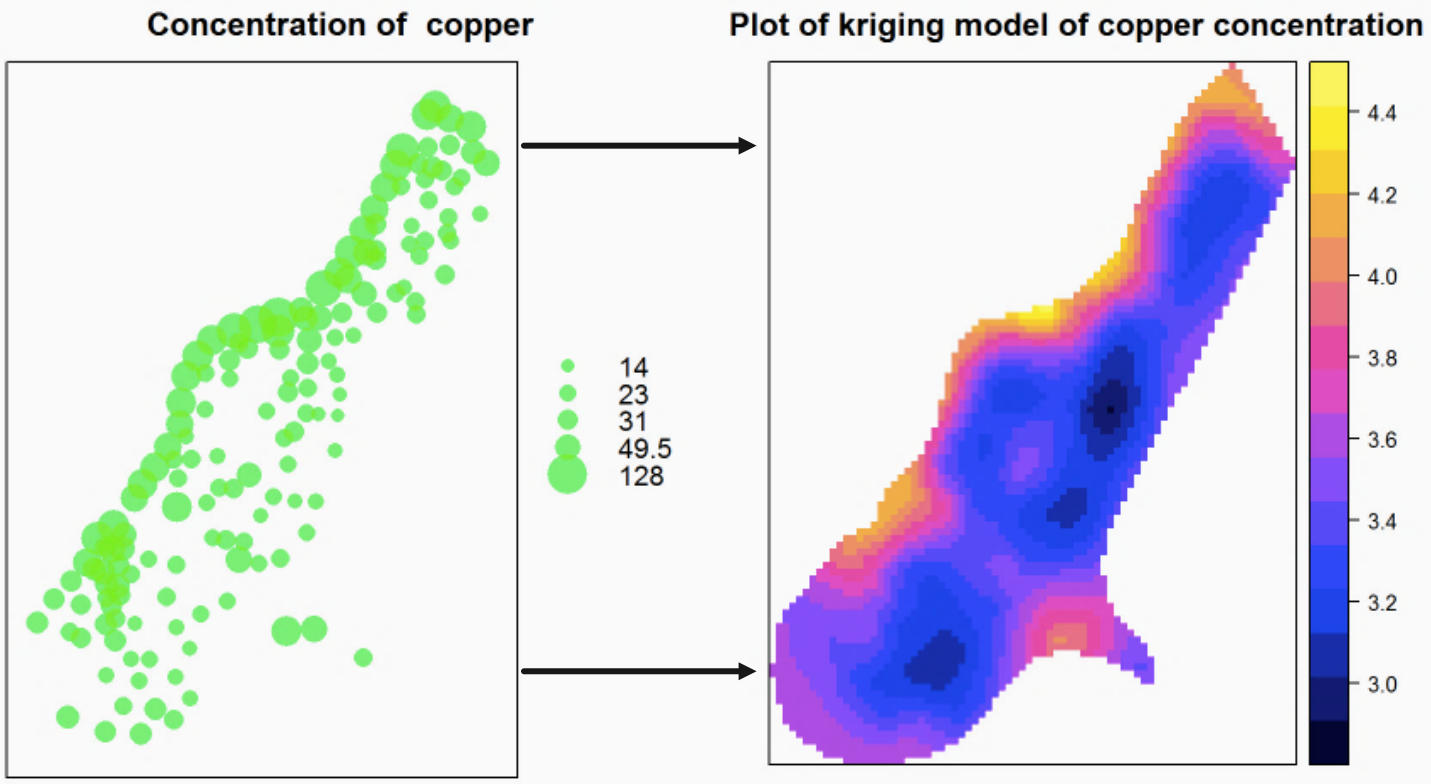
To reiterate, the process of Kriging allows us to compute all necessary components completely:

$$\underset{\boldsymbol{K^{\text{neighbors}}}}{\begin{bmatrix} 1.9 & 1.2 & \cdots & 0.005 \\ 1.2 & 1.9 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 0.005 & \cdots & \cdots & 1.9 \end{bmatrix}} \quad \underset{\boldsymbol{K^x}}{\begin{bmatrix} 0.5 \\ 0.6 \\ \vdots \\ 0.2 \end{bmatrix}} \qquad k(x,x) = 1.9 \qquad \mu = \sum_y y^i$$

$$\lambda^x = K^x K^{\text{neighbors}^{-1}}$$

$$\hat{y}(x) = \mu + \sum_{i \text{ neighbours}} \lambda_i^x \cdot (y^i - \mu)$$

$$\hat{\sigma}^2(x) = k(x,x) - \sum_i \lambda_i^x k(x^i, x)$$



To further simplify the equation, the mean μ can be removed and substitute the vectorized notation:

From:

$$\lambda^x = K^x K^{\text{neighbors}^{-1}}$$

$$\hat{\gamma}(x) = \mu + \sum_{i \text{ neighbours}} \lambda_i^x \cdot (y^i - \mu)$$

$$\hat{\sigma}^2(x) = k(x,x) - \sum_i \lambda_i^x k(x^i, x)$$

To:

$$\lambda^x = K^x K^{\text{neighbors}^{-1}}$$

$$\hat{\gamma}(x) = \sum_{i \text{ neighbours}} \lambda_i^x \cdot y^i = \lambda^x = K^x \cdot K^{\text{neighbors}^{-1}} \cdot y^{\text{neighbors}}$$

$$\hat{\sigma}^2(x) = k(x,x) - K^x \cdot K^{\text{neighbors}^{-1}} \cdot K^{xT}$$

Kriging (Part III) · Focused on the Gaussian Processes Derivation

Assuming the lables y^i come from a normal distribution N with a mean $\mu = 0$ and a covariance matrix where the top rows represent all of the neighbours with the last row representing the xy pair where a prediction is made.

Maximum Likelihood Interpetration leads to the same math:

$$\begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x^1,x^1) & k(x^1,x^2) & k(x^1,x^3) & \cdots & k(x,x^1) \\ & k(x^2,x^2) & \cdots & & k(x,x^2) \\ & & \ddots & & \vdots \\ k(x,x^1) & & & k(x,x^n) & k(x,x) \end{bmatrix}\right)$$

Covariance of the neighbors with each other
Covariance of the neighbours with x
Variance of x with itself

The above in matrix-vector notation:

$$\begin{bmatrix} y^{\text{neighbors}} \\ \\ y \end{bmatrix} \sim N\left(0, \begin{bmatrix} K^{\text{neighbours}} & K^{xT} \\ K^x & k(x,x) \end{bmatrix}\right)$$

The distribution of y given the value of all the neighboring labels is normal N with the mean and the variance:

likelihood: $p(y|y^{\text{neighbors}})$ is $N(K^x K^{\text{neighbours}^{-1}} y^{\text{neighbors}}, k(x,x) - K^x K^{\text{neighbours}^{-1}} K^{xT})$

y is estimated to be the mean: $\hat{y}(x) = K^x K^{\text{neighbours}^{-1}} y^{\text{neighbors}}$

The Variance of y is estimated: $\text{Var}(\hat{y}) = k(x,x) - K^x K^{\text{neighbours}^{-1}} K^{xT}$

Proof of the Kriging Optimization Process identical to the Gaussian Process Derivation:

From optimization:

Predictions $\hat{y}(x) = \sum_{i \text{ neighbours}} \lambda_i^x \cdot y^i = \lambda^x = K^x \cdot K^{\text{neighbors}^{-1}} \cdot y^{\text{neighbors}}$

Variance of Prediction Error $= \hat{\sigma}^2(x) = k(x,x) - K^x \cdot K^{\text{neighbors}^{-1}} \cdot K^{xT}$

Notes regarding Kriging · Gaussian Processes · Spatial Regression

- Excellent modeling tool.
- The dimensionality of the input space does not matter; just estimate of covariances $k(x,x)$.
- k can be arbitrarily complicated.
- The technique is easily adapted to use different local means.
- Vanilla Gaussian Processes scale poorly (cubically) in the number of points; it cannot be done for more than a few thousand points without having to use fancier techniques.