Reconstruction of High Dimensional Functions from Irregularly Spaced and Error Afflicted Samples by Kriging

Robert Dalitz, Eric von Lieres
e.von.lieres@fz-juelich.de
Institute of Bio- and Geosciences 1, Research Center Jülich
Wilhelm-Johnen-Str. 1, 52425 Jülich, Germany

Abstract

Interpolation is used for constructing new data points from known samples of an underlying unknown function. The majority of interpolation methods, in particular in higher dimensions, require the known data points to be placed on some equidistant grid. However, it is a common situation to have only certain measurements at irregularly spaced positions available that are additionally afflicted with measurement errors. Kriging, a class of linear least square estimators that is named after Daniel Gerhardus Krige and was first applied in geostatistics, is very well suited for such situations. The presented ordinary Kriging approach is the most popular and commonly used Kriging method because of its effectiveness and simplicity.

1 Introduction

A common interpolation situation is to have n positions $x_1, \dots, x_n \in \mathcal{D} \subset \mathbb{R}^p$ with corresponding function values or measurements $f(x_1), \dots, f(x_n) \in \mathbb{R}$ of an unknown function $f: \mathcal{D} \longrightarrow \mathbb{R}$ available. The aim is to estimate a value $f^*(\hat{x})$ of the true but unknown value $f(\hat{x})$ at a location $\hat{x}$. In the terminology of Kriging methods the function f is often called p*rocess*. According to [1] it is treated as a random process of which expected values and variances exist. Moreover its domain $\mathcal{D}$ must be compact.

The basic idea of all Kriging methods is that the underlying function $f(x)$ can be decomposed into several components. In the easiest case $f(x)$ consists of a trend component $m(x)$ and a residual component $r(x)$ such that

$$f(x) = m(x) + r(x) \tag{1}$$

in which $E[f(x)] = m(x)$ holds. Kriging methods assume that the residual component of the sought value is a weighted mean of the measurement residuals

$$r^*(\hat{x}) = \sum_{i=1}^{n} \lambda_i r(x_i) \tag{2}$$

with $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. Inserting (1) yields

$$f^*(\hat{x}) = m(\hat{x}) + \sum_{i=1}^{n} \lambda_i \big(f(x_i) - m(x_i) \big). \tag{3}$$

Because $m(x)$ is usually unknown further assumptions are necessary. At this point several options are possible (for further details see [3]). Ordinary Kriging assumes a constant but still unknown trend component $m(x) := m$ so that the above equation can be written as

$$f^*(\hat{x}) = \left(1 - \sum_{i=1}^{n} \lambda_i \right) m + \sum_{i=1}^{n} \lambda_i f(x_i). \tag{4}$$

In order to eliminate m the term within the brackets has to be zero. Therefore the assumption $\sum_i \lambda_i = 1$ must hold, which reduces equation (4) to a weighted sum of known values

$$f^*(\hat{x}) = \sum_{i=1}^{n} \lambda_i f(x_i). \tag{5}$$

For calculating the weights two more steps are required. At first the spatial variation behavior of $f(x)$ has to be modeled. Second the equation system for the weights can be set up and solved.

2 Modeling Spatial Variation

To capture the spatial variation behavior of the function $f(x)$, nearby function values of the residual component $r(x)$ are characterized. Therefore $r(x)$ is assumed as a *stationary process*. According to [3] a stationary process can be defined in two ways:

$r: \mathcal{D} \longrightarrow \mathbb{R}$ is called

Intrinsic stationary (IS) if

$E[r(x)] = \mu$ for all $x \in \mathcal{D}$ and a fixed $\mu \in \mathbb{R}$ and
$\frac{1}{2} Var[r(x+h) - r(x)] = \frac{1}{2} Var[r(h) - r(0)] =: \gamma(h)$ for all shifts h.

Second-order stationary (SOS) if

$E[r(x)] = \mu$ for all $x \in \mathcal{D}$ and a fixed $\mu \in \mathbb{R}$ and
$Cov[r(x+h), r(x)] = Cov[r(h), r(0)] =: C(h)$ for all shifts h.

$\gamma(\mathrm{h})$ and $\mathrm{C}(\mathrm{h})$ are referred to as *variogram* and *covariogram*, respectively. The definition of IS means that the variance of the difference between two measurements does not depend on the absolute position of both points, but only on their spatial distance. The same applies to SOS and the covariance. The functions are closely connected to each other because

$$\gamma(h) = C(0) - C(h) \tag{6}$$

holds which can be derived with elementary calculations. Either $\gamma(h)$ or $C(h)$ has to be determined through a fitting model $\gamma^*(h)$ or $C^*(h)$ respectively. Since the definition of IS is more general than SOS (IS can still hold in situations where $C(0)$ might not exist, which is assumed in this paper; for more details see [3]) the following derivation is based on the variogram whereas the covariogram is still used for explanation purposes.

2.1 Models

To simplify the modeling task ordinary Kriging assumes that $\gamma(h)$ and $C(h)$ only depend on the distance between two measurement locations, i.e. $\gamma(h) = \gamma(\|h\|)$ and $C(h) = C(\|h\|)$ respectively.

The existing (co-)variogram models are based on several observations. If $h \to 0$ then $\gamma(h) \to 0$ and $C(h) \to Var[r(x)] =: \sigma^2$ by definitions. And if $h \to \infty$ then $C(h) \to 0$ which means $\gamma(h) \to C(0) = \sigma^2$ due to equation (6).

Moreover (co-)variogram models have to satisfy the following condition according to [4]. For the variance of $r(\hat{x})$ the calculation

$$\begin{aligned} Var[r(\hat{x})] &\approx Var\left[\sum_{i=1}^{n} \tau_i r(x_i)\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \tau_i \tau_j Cov\left[r(x_i), r(x_j)\right] \\ &\approx \sum_{i=1}^{n}\sum_{j=1}^{n} \tau_i \tau_j C^*(h_{i,j}) \geq 0 \end{aligned} \tag{7}$$

holds for any numbers τ_i. The last inequality can be fulfilled by the fact that every variance must be positive. This is the exact definition of positive definiteness of the function $C^*(h)$, so a valid covariogram model $C^*(h)$ must be positive definite.
Using (6) in this calculation yields

$$Var[r(\hat{x})] \approx C^*(0) \sum_{i=1}^{n} \tau_i \sum_{j=1}^{n} \tau_j - \sum_{i=1}^{n}\sum_{j=1}^{n} \tau_i \tau_j \gamma^*(h_{i,j}) \geq 0. \tag{8}$$

As mentioned before there are situations where $C^*(0)$ does not exist. This term cancels out if $\sum_i \tau_i = 0$ holds so that (8) reduces to

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \tau_i \tau_j \gamma^*(h_{i,j}) \leq 0. \tag{9}$$

Thus $\gamma^*(h)$ must be conditionally negative definite.

A widely-used variogram model is the *powered exponential model*

$$\gamma^*(h) = \sigma^2 \left(1 - e^{-\left(\frac{\|h\|}{\alpha}\right)^{\beta}} \right) \tag{10}$$

with $\sigma^2 > 0, \alpha > 0$ and $\beta \in]0,2]$. Several other models have been proposed for special interpolation situations which differ regarding the origins of given data. The application of the powered exponential model is recommendable because it is relatively simple and covers a wide range of different possible behaviors of the spatial variation to be modeled. Further details can be found in [5].

If the measurements $f(x_1), \dots, f(x_n)$ were exact, a variogram model that was constructed by the considerations made so far would be sufficient. However if the given data are noisy, a modification of the variogram model is necessary.
It is assumed that the residual component $r(x)$ that was considered so far is error afflicted, now called $r_e(x)$. It decomposes into a true part $r(x)$ and an error component $e(x)$ so that

$$r_e(x) = r(x) + e(x). \tag{11}$$

The sum of two valid variogram models is also a valid model because (9) would still hold. So the chosen model, i.e. the powered exponential model can be used as variogram model for $r(x)$. Only an additional variogram model for the error component is needed.

According to [3] the variance of $e(x)$ is assumed to be constant but still unknown so that $Var[e(x)] = Cov[e(x), e(x)] =: \sigma_e^2$. This model parameter is called *nugget*. Furthermore it should be a random error which means $Cov\left[e(x_i), e\left(x_j\right)\right] = 0$ should hold for $x_i \neq x_j$. All in all this yields the covariogram model

$$C_e^*(h) = \begin{cases} \sigma_e^2, & \|h\| = 0 \\ 0, & \|h\| > 0 \end{cases}, \tag{12}$$

and by inserting this into (6) the corresponding variogram model

$$\gamma_e^*(h) = \begin{cases} 0, & \|h\| = 0 \\ \sigma_e^2, & \|h\| > 0 \end{cases} \tag{13}$$

is obtained. Hence the so called *powered exponential variogram with nugget* is the sum of (10) and (13) which is

$$\gamma^*(h) = \begin{cases} 0, & \|h\| = 0 \\ \sigma_e^2 + \sigma^2 \left(1 - e^{-\left(\frac{\|h\|}{\alpha}\right)^{\beta}} \right), & \|h\| > 0 \end{cases} \tag{14}$$

with parameters $\sigma_e^2 \geq 0, \sigma^2 > 0, \alpha > 0$ and $\beta \in]0,2]$. These parameters have to be determined in the next step.

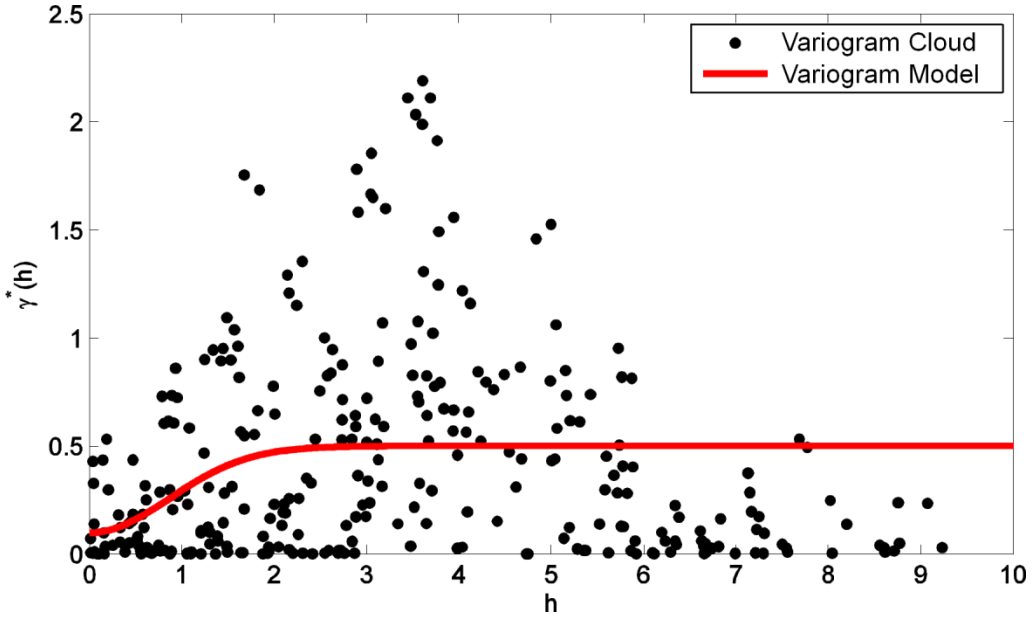

Figure 1: A powered exponential variogram model obtained by an example variogram cloud. $\gamma^*(h) = 0.0982 + 0.4026\left(1 - e^{-\left(\frac{\|h\|}{1.2274}\right)^2}\right)$ for $h > 0$

2.2 Parameter Estimation

The definition of a variogram can be transformed into a form with an expected value

$$\gamma(h) = \frac{1}{2}Var[r(x+h) - r(x)] = \frac{1}{2}E\left[\left(r(x+h) - r(x)\right)^2\right]. \qquad (15)$$

By solving (1) for $r(x)$ and inserting the result into (15) the function $r(x)$ can be eliminated. Because $m(x)$ is considered constant in ordinary Kriging this yields

$$\gamma(h) = \frac{1}{2}E\left[\left(f(x+h) - f(x)\right)^2\right]. \qquad (16)$$

Thus $y_{i,j} := \frac{1}{2}\left(f(x_i) - f(x_j)\right)^2$ can be used as observed estimates for $\gamma(h_{i,j})$ with $h_{i,j} := \|x_i - x_j\|_2$. These values form the *variogram cloud*

$$\left(h_{i,j}, y_{i,j}\right) = \left(\|x_i - x_j\|_2, \frac{1}{2}\left(f(x_i) - f(x_j)\right)^2\right) \in \mathbb{R}^2, \qquad i < j. \qquad (17)$$

The variogram model parameters can now be obtained by performing a fitting algorithm, e.g. least squares, that fits the variogram model (14) to the variogram cloud (17). This is illustrated by figure 1.

3 Equation System

The weights $\lambda_1, \dots, \lambda_n$ in (5) are determined such that the error variance is minimized as described in [6]. Therefore the error variance can be rearranged to

$$\begin{aligned} &Var\big(f(\hat{x}) - f^*(\hat{x})\big) \\ &\quad = \sum_{i=1}^{n} \lambda_i E\left[\big(f(\hat{x}) - f(x_i)\big)^2\right] \\ &\quad - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j E\left[\Big(f(x_i) - f\big(x_j\big)\Big)^2\right] \\ &\quad \approx 2\sum_{i=1}^{n} \lambda_i \gamma^*(\hat{x} - x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma^*\big(x_i - x_j\big) \\ &\quad = : V(\boldsymbol{\lambda}). \end{aligned} \tag{18}$$

To minimize $V(\boldsymbol{\lambda})$ subject to the constraint $\sum_i \lambda_i = 1$, which the weights have to fulfill, the function

$$W(\boldsymbol{\lambda}, \theta) := V(\boldsymbol{\lambda}) - \theta\left(\sum_{i=1}^{n} \lambda_i - 1\right) \tag{19}$$

has to be minimized with a Langrange multiplier θ. Therefore the derivatives

$$\frac{dW}{d\lambda_i} = 2\gamma^*(\hat{x} - x_i) - 2\sum_{j=1}^{n} \lambda_j \gamma^*\big(x_i - x_j\big) - \theta, \quad \frac{dW}{d\theta} = -\sum_{j=1}^{n} \lambda_j + 1 \tag{20}$$

have to be zero. That leads to

$$\sum_{j=1}^{n} \lambda_j \gamma^*\big(x_i - x_j\big) + \frac{\theta}{2} = \gamma^*(\hat{x} - x_i), \quad \sum_{i=1}^{n} \lambda_i = 1 \tag{21}$$

which must hold for every x_i with $i = 1, \dots, n$. These $n + 1$ equations arranged in an equation system give

$$\underbrace{\begin{bmatrix} \gamma^*(x_1 - x_1) & \cdots & \gamma^*(x_1 - x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma^*(x_n - x_1) & \cdots & \gamma^*(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}}_{=: \bar{\Gamma}} \underbrace{\begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \frac{\theta}{2} \end{bmatrix}}_{=: \bar{\lambda}} = \underbrace{\begin{bmatrix} \gamma^*(\hat{x} - x_1) \\ \vdots \\ \gamma^*(\hat{x} - x_n) \\ 1 \end{bmatrix}}_{=: \bar{\gamma}} \tag{22}$$

by which the weights are uniquely determined. Afterwards an approximate value of $f(\hat{x})$ can be obtained by inserting $\lambda_1, \dots, \lambda_n$ into (5), which yields

$$f^*(\hat{x}) = \bar{y}^T \bar{\lambda} = \bar{y}^T \bar{\Gamma}^{-1} \bar{\gamma}(\hat{x}) \tag{23}$$

with $\bar{y}_O := [f(x_1), \dots, f(x_n), 0]^T$. This formula can be used for every location where a value should be estimated because $\bar{\Gamma}$ is independent of $\hat{x}$. Figure 2 shows an example of an approximating function computed by ordinary Kriging.

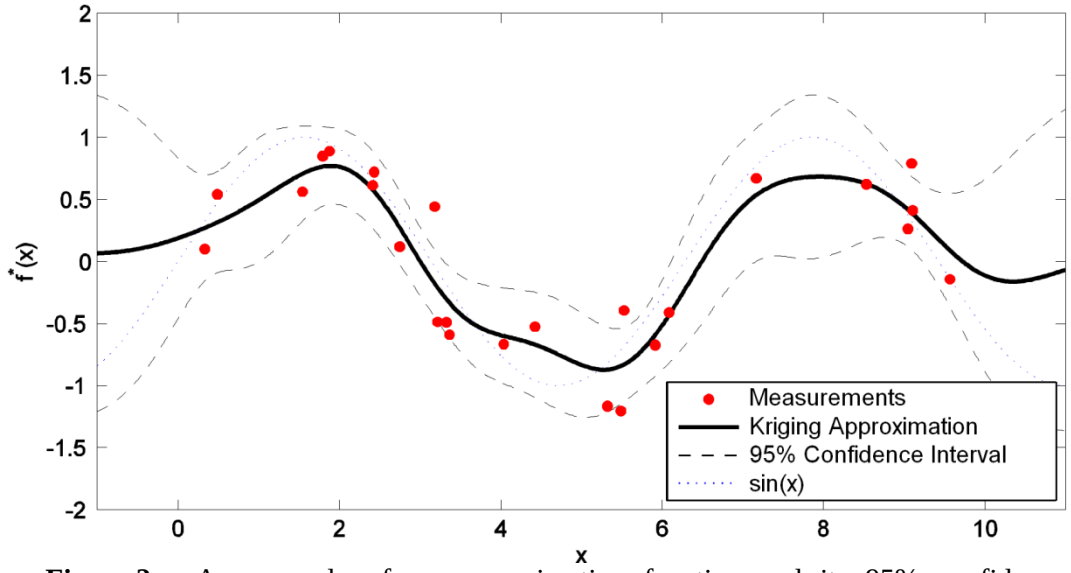

Figure 2: An example of an approximating function and its 95% confidence interval computed by Kriging. The measurements were generated by $\sin(x)$ and a random error within the interval $\left[-\frac{1}{2},\frac{1}{2}\right]$. The corresponding variogram model is shown in Figure 1.

3.1 Error Estimation

Kriging methods offer the possibility to estimate the error variance of the approximation. It can be easily obtained from $V(\boldsymbol{\lambda})$ in equation (18) as follows:

$$\begin{aligned} V(\hat{x}) &= 2\sum_{i=1}^{n}\lambda_i\gamma^*(\hat{x}-x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma^*(x_i-x_j) \\ &= \sum_{i=1}^{n}\lambda_i\gamma^*(\hat{x}-x_i) + \frac{\theta}{2} = \bar{\gamma}^T\bar{\lambda} = \bar{\gamma}(\hat{x})^T\bar{\Gamma}^{-1}\bar{\gamma}(\hat{x}) \end{aligned} \tag{24}$$

This value is also called *Kriging variance*. When weights have already been computed, the effort for calculating this variance is computationally cheap because only one more matrix-vector multiplication is necessary.

According to [3] the Kriging variance can be used to calculate a confidence interval as

$$\left[f^*(\hat{x}) - z\sqrt{V(\hat{x})}\,, f^*(\hat{x}) + z\sqrt{V(\hat{x})}\right]. \tag{25}$$

This means that the true but unknown value $f(\hat{x})$ is within the confidence interval with a probability depending on z. If a p% confidence interval is desired z has to be the $\left(\frac{1}{2}+\frac{p}{200}\right)$-quantile of the standard normal distribution. E.g. $z = 1.96$ yields a 95% confidence interval. A confidence interval is also shown in Figure 2.

4 Conclusions

Ordinary Kriging is a method that is suitable for a wide range of interpolation problems, in particular where derivatives are not available and further function evaluations are expensive and/or time-consuming. Moreover, the error variance that can be easily calculated, is often very important. Ordinary Kriging is the basis for many other Kriging variants that modify certain ideas to meet more specific requirements [7].

5 References

[1] Jarre, Florian; Stoer, Josef: *Optimierung*. Springer, 2003.

[2] Chilès, Jean-Paul; Delfiner, Pierre: *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, 1999.

[3] Sherman, Michael: *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. John Wiley & Sons, 2011.

[4] Goovaerts, Pierre: *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 1997.

[5] Diggle, Peter J.; Ribeiro, Paulo Justiniano: *Model-based Geostatistics*. Springer, 2007.

[6] Dalitz, Robert: *Versuchsplanung und Datenauswertung für Hochdurchsatzexperimente mit Genetischen Algorithmen und Kriging-Verfahren*. Master's thesis, Heinrich-Heine-Universität Düsseldorf, 2011.

[7] Cressie, Noel A. C.: *Statistics for Spatial Data*. John Wiley & Sons, revised edition, 1993.