Interpolation, Kriging, Gaussian Processes

CEE 629. System Identification

Department of Civil and Environmental Engineering
Duke University

Henri P. Gavin and Suraj Khanal
Fall 2021

1 Introduction

Given a sample of m measurements, $y(x_1)$, ..., $y(x_m)$, at m sample points, $x_1, \cdots, x_m$ ($x_i \in R^n, y \in R^1, n \geq 1$), methods of multi-dimensional interpolation provide estimates of the value $\hat{y}(x_o)$ at the interpolation point x_o. The sample of measurement locations $(x_1, \cdots, x_m)$ may be randomly distributed or uniformly spaced. In interpolated point $\hat{y}(x_o)$ may be computed as a weighted average of the measurements $y_1, \cdots, y_m$ ($y(x_1)$, ..., $y(x_m)$).

$$\hat{y}(x_o) = \sum_{i=1}^{m} w_i(x_o)\ y(x_i)\ , \tag{1}$$

in which the weights, w_i, depend on the interpolation location x_o, and the weights sum to unity,

$$\sum_{i=1}^{m} w_i = 1\ . \tag{2}$$

2 General multi-dimensional interpolation

Interpolation methods are distinguished by the means of determining values for the weights, w_i, in the interpolation equation (1)

$$\hat{y}(x_o) = \sum_{i=1}^{m} w_i(x_o)\ y(x_i)\ .$$

In general, the closer x_i is to x_o, the more y_i should influence the interpolated estimate $\hat{y}_o$. So the weights $w_i(x_o)$ should be smaller for points x_i that are close to x_o, and small for x_i values far from x_o. For example, the weights $w_i(x_o)$ can be specified to decrease with the (Euclidean) distance d_{oi} between x_o and x_i,

$$d_{oi} = |x_o - x_i| = \sqrt{(x_o - x_i)^\mathsf{T}(x_o - x_i)}\ . \tag{3}$$

Examples of so-called "inverse distance weighted" (IDW) interpolation weights include:

$$w_i \propto \frac{1}{1 + (d_{oi}/\alpha)^q}\ , \tag{4}$$

and

$$w_i \propto \left(\frac{d_{\max} - d_{oi}}{d_{\max}\ d_{oi}}\right)^q\ , \tag{5}$$

where $d_{\max}$ is defined as $\max_{i=1,\cdots,n}(d_{oi})$, $\alpha > 0$, and $q > 0$.

Notes:

- The analyst has a choice of hyper-parameters $d_{\max}$, $q > 0$, and $\alpha > 0$.
- These weights do not depend on the characteristics of the data being interpolated. They depend only on the distribution of the sample points x_i with respect to the interpolation point x_o.
- The weights in equation (4) and (5) are *not* normalized ($\sum w_i \neq 1$), so for these cases, the interpolation equation should include a normalization factor

$$\hat{y}(x_o) = \sum_{i=1}^{m} w_i(x_o)\ y(x_i)\ / \sum_{i=1}^{m} w_i(x_o)\ . \tag{6}$$

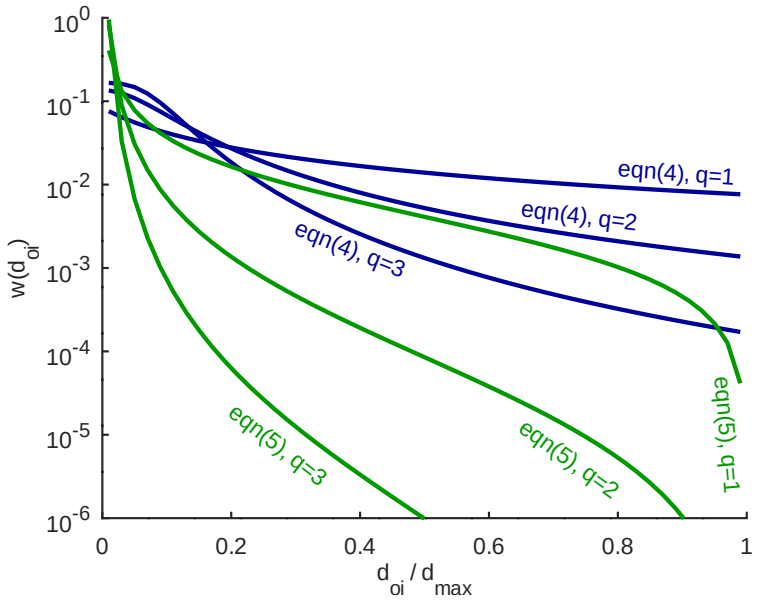

Figure 1. The effect of the exponent q on IDW weights in equations (4) and (5), $\alpha = 0.1 d_{\max}$

3 Optimal least-squares interpolation ("simple Kriging")

In Kriging interpolation, the weights in the interpolation equation (1)

$$\hat{y}(x_o) = \sum_{i=1}^{m} w_i(x_o)\ y(x_i)$$

minimize a squared error

$$e = \left(y(x_o) - \sum_{i=1}^{m} w_i(x_o)\ y(x_i) \right)^2 \tag{7}$$

such that $\sum w_i = 1$.

The error is not summed over a number of data points, as is done with least-squares curve-fitting. Least squares curve fitting is appropriately applied to data sets in which:

- one or more measurements are recorded at several values of the independent variables;
- the measurement error is significant;
- a goal is to fit a curve *through* a cloud of measured data points, but not through any of the individual points in particular;
- another goal may be to identify the parameters of some physically-motivated model; and
- the measurements are known in advance to vary according to some known function of the independent variables and the model parameters.

Least squares interpolation, on the other hand, is appropriately applied to data sets in which:

- only one measurement is taken at each value of the independent variables;
- the measurement error is almost insignificant;
- a goal is to fit a curve that passes exactly through each of the measured data points;
- the parameter values (the weights) are not needed to provide much, if any, insight into the system being measured; and
- the fashion in which the measurements vary as a function of the independent variables is not known in advance, and can be quite irregular, as in mountainous terrain.

Curve-fitting is often applied to parameter estimation and system identification problems. Interpolation is often applied to meta-modeling problems.

3.1 Three identities

Methods of Kriging interpolation make use of three identities derived below.

The squared error function may be expanded as follows.

$$\begin{aligned} e &= y_o^2 - 2y_o \sum_i w_i y_i + \sum_j w_j y_j \sum_i w_i y_i && (8)\\ &= y_o^2 - 2y_o \sum_i w_i y_i + \sum_i \sum_j w_i y_i \; w_j y_j && (9) \end{aligned}$$

$(y_o = y(x_o), w_i = w_i(x_o), y_i = y(x_i))$.

The following two identities are valid only for $\sum w_i = 1$.

$$\begin{aligned} 2\sum_i w_i \frac{1}{2}(y_o - y_i)^2 &= \sum_i w_i (y_o^2 - 2y_o y_i + y_i^2) && (10)\\ &= y_o^2 \sum_i w_i - 2y_o \sum_i w_i y_i + \sum_i w_i y_i^2 && (11)\\ &= y_o^2 - 2y_o \sum_i w_i y_i + \sum_i w_i y_i^2 && (12) \end{aligned}$$

$$\begin{aligned} \sum_i \sum_j w_i w_j \frac{1}{2}(y_i - y_j)^2 &= \frac{1}{2}\sum_i \sum_j w_i w_j (y_i^2 - 2y_i y_j + y_j^2) && (13)\\ &= \frac{1}{2}\sum_i \sum_j w_i w_j y_i^2 - \sum_i \sum_j w_i w_j y_i y_j + \frac{1}{2}\sum_i \sum_j w_i w_j y_j^2 && (14)\\ &= \frac{1}{2}\sum_j w_j \sum_i w_i y_i^2 - \sum_i \sum_j w_i w_j y_i y_j + \frac{1}{2}\sum_i w_i \sum_j w_j y_j^2 && (15)\\ &= \frac{1}{2}\sum_i w_i y_i^2 - \sum_i \sum_j w_i w_j y_i y_j + \frac{1}{2}\sum_j w_j y_j^2 && (16)\\ &= \sum_i w_i y_i^2 - \sum_i \sum_j w_i w_j y_i y_j && (17) \end{aligned}$$

So the squared error criterion can be expressed in terms of $(y_o - y_i)^2$ and $(y_i - y_j)^2$.

$$\begin{aligned} e &= y_o^2 - 2y_o \sum_i w_i y_i + \sum_i \sum_j w_i w_j y_i y_j\\ &= y_o^2 - 2y_o \sum_i w_i y_i + \sum_i w_i y_i^2 - \sum_i w_i y_i^2 + \sum_i \sum_j w_i w_j y_i y_j\\ &= 2\sum_i w_i \frac{1}{2}(y_o - y_i)^2 - \sum_i \sum_j w_i w_j \frac{1}{2}(y_i - y_j)^2 \; . && (18) \end{aligned}$$

3.2 The "semivariogram"

In the sense of statistical averages, the squared differences $(y_i - y_j)^2$ tend to increase with the magnitude d_{ij} of the separation between x_i and x_j,

$$\begin{aligned} d_{ij} &= |x_i - x_j| && (19)\\ &= \left[\sum_k (x_{ki} - x_{kj})^2\right]^{1/2} && (20)\\ &= \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ni} - x_{nj})^2} && (21) \end{aligned}$$

The greater the distance between x_i and x_j, the greater the expected difference between the measured observations $y(x_i)$ and $y(x_j)$. Define γ_{ij} as half of the statistical expectation of the squared difference between $y(x_i)$ and $y(x_j)$

$$\begin{aligned} \gamma_{ij} &= \frac{1}{2}\mathbb{E}\left[(y(x_i) - y(x_j))^2\right] && (22)\\ &= \frac{1}{2}\mathbb{V}\left[y(x_i) - y(x_j)\right] && (23) \end{aligned}$$

The function γ_{ij} is termed the *semivariogram* of its random field. If the random field is stationary and isotropic then γ_{ij} depends only on the separation distance $d_{ij} = |x_i - x_j|$, and increases with d_{ij}. The manner in which γ_{ij} increases with d_{ij} depends on the nature of the data. A number of semivariogram approximation functions, $\hat{\gamma}(d)$, have been proposed:

Kriging semivariogram approximation functions

Gaussian	$\hat{\gamma}(d) = c_1\delta(d) + c_2\{1 - \exp\left[-\frac{1}{2}(d/c_3)^2\right]\}$	
exponential	$\hat{\gamma}(d) = c_1 + c_2\{1 - \exp[-d/c_3]\}$	
sinc	$\hat{\gamma}(d) = c_1 + c_2\{1 - \sin(d/c_3)/(d/c_3)\}$	
power-law	$\hat{\gamma}(d) = c_1 + c_2(d/d_{\max})^{c_3}$	$0 \le c_3$, c_3 not even
spherical	$\hat{\gamma}(d) = c_1 + c_2\{\frac{3}{2}(\frac{d}{c_3}) - \frac{1}{2}(\frac{d}{c_3})^3\}$	$0 < d < c_3$
	$\hat{\gamma}(d) = c_1 + c_2$	$c_3 \le h$
rational quadratic	$\hat{\gamma}(d) = c_1 + c_2(d/c_3)^2/(1 + (d/c_3)^2)$	
linear	$\hat{\gamma}(d) = c_1\delta(d) + c_3 h$	

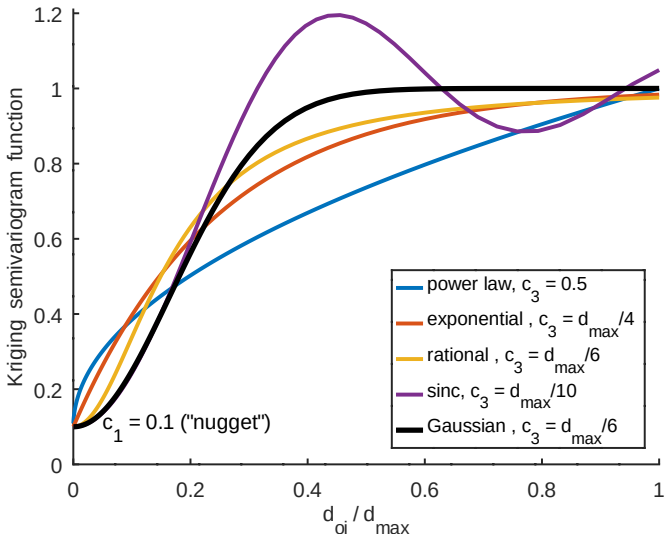

Figure 2. Kriging semivariogram approximation functions for $c_1 = 0.1$, $c_2 = 1.0$, and c_3 (shown in the legend).

Notes:

- The matrix of semivariogram values $[\hat{\gamma}(d_{ij})]$ must be positive definite, $\sum_i \sum_j w_i \, \hat{\gamma}(d_{ij}) \, w_j > 0 \ , \ \forall \ |w| \neq 0$.
- $\hat{\gamma}(0) = 0$, always, by definition. The diagonal of the matrix of semivariogram values $[\hat{\gamma}(d_{ij})]$ must be zero.
- The choice of the semivariogram function depends on the nature of the data being interpolated.
- For any set of realistic data, variability (uncertainty) in the estimate of the semivariogram becomes quite large as d becomes large. It is more important for the semivariogram function to fit the semivariogram data for small values of d than for large values of d. A "power-law" semivariogram function can work well in this regard. (opinion)
- As will be explored in the next section, the semivariogram is related to the covariance of the data, Least squares interpolation when formulated in terms of covariance functions is called Gaussian process modeling. The semivariogram is defined in equation (22) as

$$\gamma(d) = \frac{1}{2}\mathbb{E}\left[(y(x) - y(x+d))^2\right] \ , \tag{24}$$

wherein $d \in \mathbb{R}^n$. The covariance function is defined as

$$V(d) = \mathbb{E}\left[y(x)y(x+d)\right] \ . \tag{25}$$

The semivariogram and the covariance are related as follows

$$\begin{aligned} 2\gamma(d) &= \mathbb{E}\left[(y(x)^2\right] - 2\mathbb{E}[y(x)y(x+d)] + \mathbb{E}\left[y(x+d)^2\right] && (26)\\ &= \mathbb{E}[y(x)] - 2V(d) + \mathbb{V}[y(x+d)] && (27)\\ &= 2(V(0) - V(d)) && (28)\\ \gamma(d) &= \mathbb{E}\left[y^2\right] - V(d) \ , && (29) \end{aligned}$$

assuming that the variance of $y(x)$ is independent of x and that the mean of y is zero. In principle, then, if $V(d)$ is always positive-valued and tends toward zero with increasing d, then $\gamma(d)$ should be positive valued and should asymptotically approach $\mathbb{E}[y^2]$ with increasing d. The numerical example in this document illustrates problems for which $\gamma(d) \to \infty$, can, however, result in good Kriging performance. This observation begs for deeper investigation.

3.3 Least-squares optimal interpolation weights

Recall the interpolation equation (1),

$$\hat{y}(x_o) = \sum_{i=1}^{m} w_i(x_o)\ y(x_i) \ ,$$

and the squared error criterion (7),

$$e = \left(y(x_o) - \sum_{i=1}^{m} w_i(x_o)\ y(x_i) \right)^2$$

We have previously seen (equation (18)) that the squared error can be expressed as

$$e = 2\sum_i w_i \frac{1}{2}(y_o - y_i)^2 - \sum_i \sum_j w_i w_j \frac{1}{2}(y_i - y_j)^2$$

In Kriging, the expected value of $\frac{1}{2}(y(x_i) - y(x_j))^2$ is approximated by an analytic semivariogram function $\hat{\gamma}(d_{ij})$ that depends only on the separation distance $d_{ij} = |x_i - x_j|$. Using the semivariogram approximation, the error function can be written

$$e = 2\sum_i w_i \hat{\gamma}_{oi} - \sum_i \sum_j w_i \hat{\gamma}_{ij} w_j \ . \tag{30}$$

($\hat{\gamma}_{oi} = \hat{\gamma}(d_{oi})$, $\hat{\gamma}_{ij} = \hat{\gamma}(d_{ij})$). To find the weights, we minimize e such that the equality constraint $\sum w_i = 1$ is satisfied via a Lagrange multiplier. Adjoining the constraint to the squared error,

$$e_{\text{adj}} = 2\sum_i w_i \hat{\gamma}_{oi} - \sum_i \sum_j w_i \hat{\gamma}_{ij} w_j - 2\mu \left(\sum_i w_i - 1 \right) \ , \tag{31}$$

where (-2μ) is the Lagrange multiplier. In order to enforce the constraint $\sum_i w_i = 1$ while minimizing e,

$$\frac{\partial e_{\rm adj}}{\partial w_i} = 0 : \qquad \hat{\gamma}_{oi} - \sum_j \hat{\gamma}_{ij} w_j - \mu = 0 \ , \tag{32}$$

$$\frac{\partial e_{\rm adj}}{\partial \mu} = 0 : \qquad \sum_i w_i - 1 = 0 \ , \tag{33}$$

which leads to the normal equations

$$\begin{bmatrix} \hat{\gamma}_{11} & \hat{\gamma}_{12} & \cdots & \hat{\gamma}_{1m} & 1 \\ \hat{\gamma}_{21} & \hat{\gamma}_{22} & \cdots & \hat{\gamma}_{2m} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\gamma}_{m1} & \hat{\gamma}_{m2} & \cdots & \hat{\gamma}_{mm} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \\ \mu \end{bmatrix} = \begin{bmatrix} \hat{\gamma}_{o1} \\ \hat{\gamma}_{o2} \\ \vdots \\ \hat{\gamma}_{om} \\ 1 \end{bmatrix} \ , \tag{34}$$

or $\mathbf{Gw} = \mathbf{g}$. As in least-squares curve fitting, the matrix $\mathbf{G}$ depends only on the measured data through the semivariogram and could be computed without relying upon an approximation $\hat{\gamma}(d_{ij})$ for γ_{ij}. The values $\gamma_{oi} = \frac{1}{2}E[(y(x_o) - y(x_i))^2]$ depend on the unknown interpolation point value $y(x_o)$, and *cannot* be evaluated from the data alone. Equation (34) results from applying the approximation $\gamma_{ij} \approx \hat{\gamma}(d_{ij})$ to both sides of the equation. (Recall that $\hat{\gamma}(d_{ij})$ is a presumed function of the distances between x_i and x_j, $d_{ij} = |x_i - x_j|$.)

Solving this set of "$m+1$" equations for "$m+1$" unknowns results in the optimal values of the weights w_i and the Lagrange multiplier μ.

Notes:

- The weights $w_i(x_o)$ depend on the nature of the data being interpolated through the choice of the semivariogram function.
- One matrix inversion and one matrix multiplication are required for all interpolation points.
- For many problems the linear semivariogram function $\hat{\gamma}(d_{ij}) = d_{ij}$ is a good starting point and can ultimately work well for many Kriging problems (opinion).

Exercise:
Show that values of the constants c_1 and c_2 in the function $\hat{\gamma}(d_{ij}) = c_1 + c_2 d_{ij}$ do not affect the values of the optimal weights w_i.

3.4 Error analysis

The squared error criterion (equation (30)) enables a systematic error analysis.

$$\begin{aligned} e &= 2\sum_i w_i\hat{\gamma}_{oi} - \sum_i\sum_j w_i\hat{\gamma}_{ij}w_j \\ &= 2\mathbf{w}^\mathsf{T}\mathbf{g} - \mathbf{w}^\mathsf{T}\mathbf{G}\mathbf{w} - \mu \qquad (35) \\ &= 2\mathbf{w}^\mathsf{T}\mathbf{g} - \mathbf{w}^\mathsf{T}\mathbf{g} - \mu \qquad (36) \\ &= \mathbf{w}^\mathsf{T}\mathbf{g} - \mu \qquad (37) \\ &= \mathbf{g}^\mathsf{T}\mathbf{G}^{-1}\mathbf{g} - \mu > 0 \qquad (38) \end{aligned}$$

So the standard error of the Kriging interpolation at x_o may be readily computed as

$$\sigma_{\hat{y}}(x_o) = \sqrt{\sum_{i=1}^{m} w_i(x_o)\ \hat{\gamma}(|x_o - x_i|)} \qquad (39)$$

Note that while the optimal weights are not sensitive to scaling of the semivariogram function, the standard error does depend on such scaling.

4 Gaussian processes

A *Gaussian Process* is the convolution of Gaussian white noise through a Gaussian kernel function. Gaussian Process data can be interpoled by the expectation of the Gaussian Process conditional upon the set of observed data.

Given a sample of m measurements, $y(x_1)$, ..., $y(x_m)$, at m sample points, $x_1, \cdots, x_m$ ($x_i \in \mathbb{R}^n, y \in \mathbb{R}^1, n \geq 1$), a Gaussian Process model provides the expected value and the variance of the random variable $Y(x_o)$ at the interpolation point x_o. Assuming the distributions of observed data values $Y(x_i), ..., Y(x_m)$ and the distribution of the random value $Y(x_o)$ at the interpolation point x_o are joint Gaussian, with a *presumed* or *fitted* function $\hat{m}(x_i)$ to approximate the mean trend of the data $\mu(x)$, and a *presumed* function $\hat{\kappa}(x_i, x_j)$ to approximate the covariance of the data $\kappa(x_i, x_j) = \mathbb{E}[(x_i - \mu(x_i))(x_j - \mu(x_j))]$ this joint Gaussian distribution may be expressed as

$$\begin{bmatrix} Y(x_1) \\ \vdots \\ Y(x_m) \\ Y(x_o) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_m) \end{bmatrix} , \begin{bmatrix} \hat{\kappa}(x_1,x_1) & \cdots & \hat{\kappa}(x_1,x_m) \\ \vdots & \ddots & \vdots \\ \hat{\kappa}(x_m,x_1) & \cdots & \hat{\kappa}(x_m,x_m) \end{bmatrix} \begin{bmatrix} \hat{\kappa}(x_1,x_o) \\ \vdots \\ \hat{\kappa}(x_m,x_o) \end{bmatrix} \right) \tag{40}$$
$$\begin{bmatrix} \hat{m}(x_o) \end{bmatrix} \quad \begin{bmatrix} \hat{\kappa}(x_o,x_1) & \cdots & \hat{\kappa}(x_o,x_m) \end{bmatrix} \quad \hat{\kappa}(x_o,x_o)$$

$$\begin{bmatrix} Y(x) \\ Y(x_o) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} [\hat{m}] \\ \hat{m}_o \end{bmatrix} , \begin{bmatrix} \left[\hat{K}\right] & \left[\hat{K}_o\right] \\ \left[\hat{K}_o\right]^\mathsf{T} & \hat{K}_{oo} \end{bmatrix} \right) \tag{41}$$

From Bayes' Theorem, the probability of y_o conditional upon on the data y is

$$p(y_o|y) = \frac{p(y|y_o)\ p(y_o)}{p(y)} \tag{42}$$

in which, for joint Gaussian distributions,

$$Y(x_o) \sim \mathcal{N}\left(\hat{m}_o, \hat{K}_{oo}\right) \tag{43}$$
$$p(y_o|y) = \frac{p(y, y_o)}{p(y)} \tag{44}$$
$$= \mathcal{N}\left(\hat{m}_o + \hat{K}_o^\mathsf{T}\ \hat{K}^{-1}(y - \hat{m})\ ,\ \hat{K}_{oo} - \hat{K}_o^\mathsf{T}\ \hat{K}^{-1}\ \hat{K}_o\right) \tag{45}$$

So the expected value of the uncertain point Y_o, conditional upon the data y is,

$$\mathbb{E}[Y_o|y] = \hat{m}_o + \hat{K}_o^\mathsf{T}\hat{K}^{-1}\ (y - \hat{m}) \tag{46}$$

Gaussian Process weighting vectors corresponding to equation (1) are $w = \hat{K}_o^\mathsf{T}\hat{K}^{-1}$. The variance of Y_o conditional upon the data y is

$$\mathbb{V}[Y_o|y] = \hat{K}_{oo} - \hat{K}_o^\mathsf{T}\ \hat{K}^{-1}\ \hat{K}_o \tag{47}$$

Relationships for the conditional distribution of a joint Gaussian distribution are derived in section 3.3 of reference 5, and in Appendices A.2 and A.3 of reference 6, below.

The covariance function $\hat{\kappa}(x_i, x_j)$ has the properties of a *kernel*, meaning that the *presumed* covariance matrix must be positive definite. A kernel function consistent with the multi-variate Gaussian distribution is the squared exponential of the Euclidean distance between x_i and x_j, $d = ||x_i - x_j||$.

$$\hat{\kappa}(x_i, x_j) = \hat{\kappa}(||x_i - x_j||) = \hat{\kappa}(d) = \sigma^2 \ \exp\left[-\frac{1}{2}\left(\frac{d}{s}\right)^2\right] + \beta^2 \ \delta(d) \tag{48}$$

The hyper-parameter σ^2 is the variance of the data $[y_1, ...y_m]$, the hyper-parameter s is the length scale of the variability in the data, the hyper-parameter β accounts for imprecise measurements for which the expectation $\mathbb{E}((Y(x_i) - Y(x_i))^2)$ is non-zero, and $\delta(\cdot)$ is the Dirac delta function.

Gaussian Process covariance approximation functions

Gaussian	$\hat{\kappa}(d) = c_1\delta(d) + c_2 \exp\left[-\frac{1}{2}(d/c_3)^2\right]$	$c_1, c_2, c_3 > 0$
exponential	$\hat{\kappa}(d) = c_1\delta(d) + c_2 \exp[-d/c_3]$	$c_1, c_2, c_3 > 0$
sinc	$\hat{\kappa}(d) = c_1\delta(d) + c_2 \sin(d/c_3)/(d/c_3)$	$c_1, c_2, c_3 > 0$
power-law	$\hat{\kappa}(d) = c_1\delta(d) + c_2(1 + d/d_{\max})^{-2/c_3}$	$c_1, c_2 > 0, c_3 < 0$
spherical	$\hat{\kappa}(d) = c_1\delta(d) + c_2\{1 - \frac{3}{2}(\frac{d}{c_3}) + \frac{1}{2}(\frac{d}{c_3})^3\}$, if $d < c_3$, 0 if $d > c_3$	$c_1, c_2, c_3 > 0$
rational quadratic	$\hat{\kappa}(d) = c_1\delta(d) + \max[0 \ , \ c_2\{1 - (d/c_3)^2/(1 + (d/c_3)^2)\}$	$c_1, c_2, c_3 > 0$
linear	$\hat{\kappa}(d) = \max[0 \ , \ c_1\delta(d) + c_2(1 - c_3 h)] + \delta(d)$	$c_1, c_2 > 0$

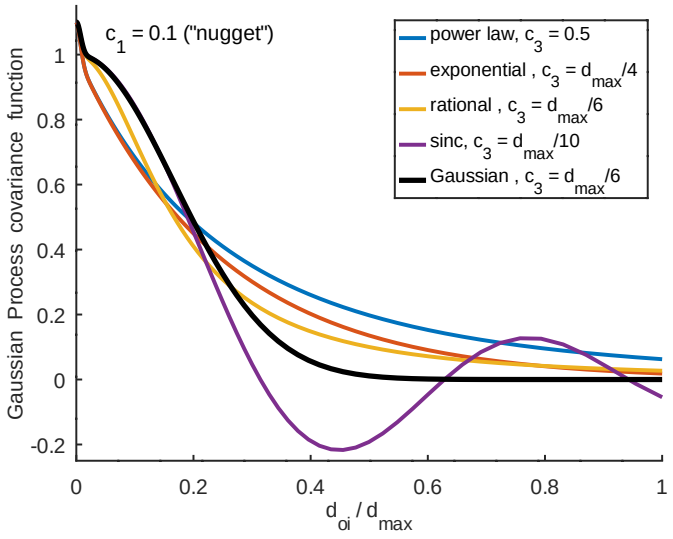

Figure 3. Gaussian Process covariance approximation functions for $c_1 = 0.1$, $c_2 = 0.9$, and c_3 (shown in the legend).

A Gaussian process based upon a Gaussian covariance function is a Gaussian mixture. Defining $\alpha \equiv \hat{K}^{-1} y$

$$\mathbb{E}[Y_o|y] = \hat{m}_o + \sum_{i=1}^{m} \alpha_i \hat{K}_{oi} \tag{49}$$

$$= \hat{m}_o + \sum_{i=1}^{m} \alpha_i \ \sigma^2 \ \exp\left[-\frac{1}{2s^2}||x_o - x_i||^2\right] \tag{50}$$

5 Numerical examples

In this section the relative accuracy of the Gaussian Process, Kriging, and inverse distance weighted interpolations are assessed by their performance in reconstructing the following function in $\mathbb{R}^2$:

$$y(x_1, x_2; L) = \sin(\pi x_1/L)\ \cos(\pi x_2/L) - 0.2 x_1 x_2 + N\ , \tag{51}$$

from samples of this function with and without observation errors. The single subscripts in equation (51) indicate the variable in $\mathbb{R}^2$. In the following, a double subscript indicates which observation of which variable is involved, i.,e., x_{kj} is the k^{th} observation of variable x_j, as in equation (21). The parameter L indicates the length scale of variability in the surface $y(x)$. The domain considered here is $-2 \le x_1, x_2 \le 2$, as shown in Figure 4. The random variable N represents measurement errors modeled as uncorrelated Gaussian noise, $N \sim \mathcal{N}(0, \sigma_N^2)$.

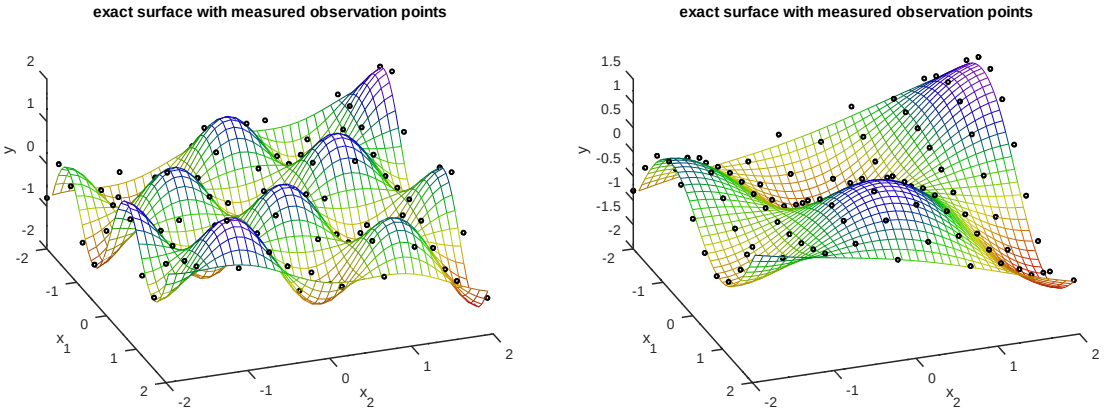

Figure 4. The example surfaces for IDW, Kriging, and Gaussian Process interpolation. (a) length scale $L = 1$, (b) length scale $L = 2$. The 105 regularly-spaced measured observation points shown in blue are barely enough to characterize the surface with short length scale ($L = 1$), but are more than adequate to characterize the surface with longer length scale ($L = 2$).

The accuracy of each interpolation method is quantified here in two ways: (a) by the root mean square (R.M.S.) of the difference between the interpolated surface, $\hat{y}(x_p)$, and the noise-free surface, given by equation (51),

$$E_{\text{rms}} = \sqrt{\frac{1}{P}\sum_{p=1}^{P}\left[y(x_p) - \hat{y}(x_p)\right]^2}\ , \tag{52}$$

where x_p are the interpolation points and P is the total number of points being interpolated; and (b) by the correlation between the interpolated surface $\hat{y}(x_p)$ and the noise-free surface, given by equation (51),

5.1 Computing semivariogram and covariance data from random samples

When samples of x are regularly spaced on a grid many pairs of sample points x_i, x_j have the same separation distance. In such cases, the squared semi-variations $\frac{1}{2}(y_i - y_j)^2$

and the covariance $(y_i)(y_j)$ of all pairs having similar separation distances may be averaged to estimate the semi-variation and the covariance corresponding to the particular separation distance. When samples of x are randomly distributed the semivariogram data may be estimated as follows:

1. Compute $d_{ij} = ||x_i - x_j||$, $v_{ij} = \frac{1}{2}(y(x_i) - y(x_j))^2$, and $c_{ij} = y(x_i)\ y(x_j)$ for all pairs of data.
2. Sort d_{ij} and the corresponding v_{ij} and c_{ij} into increasing numerical order of d_{ij}.
3. Group the remaining values of d into non-overlapping segments. For each segment, compute the average of the corresponding values of v and c, and determine the midpoint of the segment of d.
4. The averaged values of v and c and the corresponding segment-midpoints of d represent the semivariogram and the covariance of the data set.

Example semivariogram and covariance data calculated with this method, and fit with a number of approximate semivariogram and covariance approximation functions, are shown in Figure 5. The length scale of the oscillatory nature of the data is revealed by the relative maximum in the semivariogram data at $d \approx L$. The semivariogram information in Figures 5 is relatively precise up to about half of the characteristic length scale of the problem. As the correlation distance d_{ij} increases beyond the characteristic length scale of the surface, variability in the semivariogram and covariance data clearly increases.

Accurate estimation of the semivariogram and covariance information requires a number of observations (measurements), m, that is sufficient to average enough terms of $\frac{1}{2}(y(x_i) - y(x_j))^2$ and $y_i y_j$ in order to achieve an acceptably small level of variability in the semivariogram data (less than ten percent, or so), without sacrificing resolution in $||x_i - x_j||$. Because this number of measurements is often much greater than required for accurate Kriging interpolation, one may question the importance of selecting a semivariogram function that is highly representative of the data. These statements are supported by the following results.

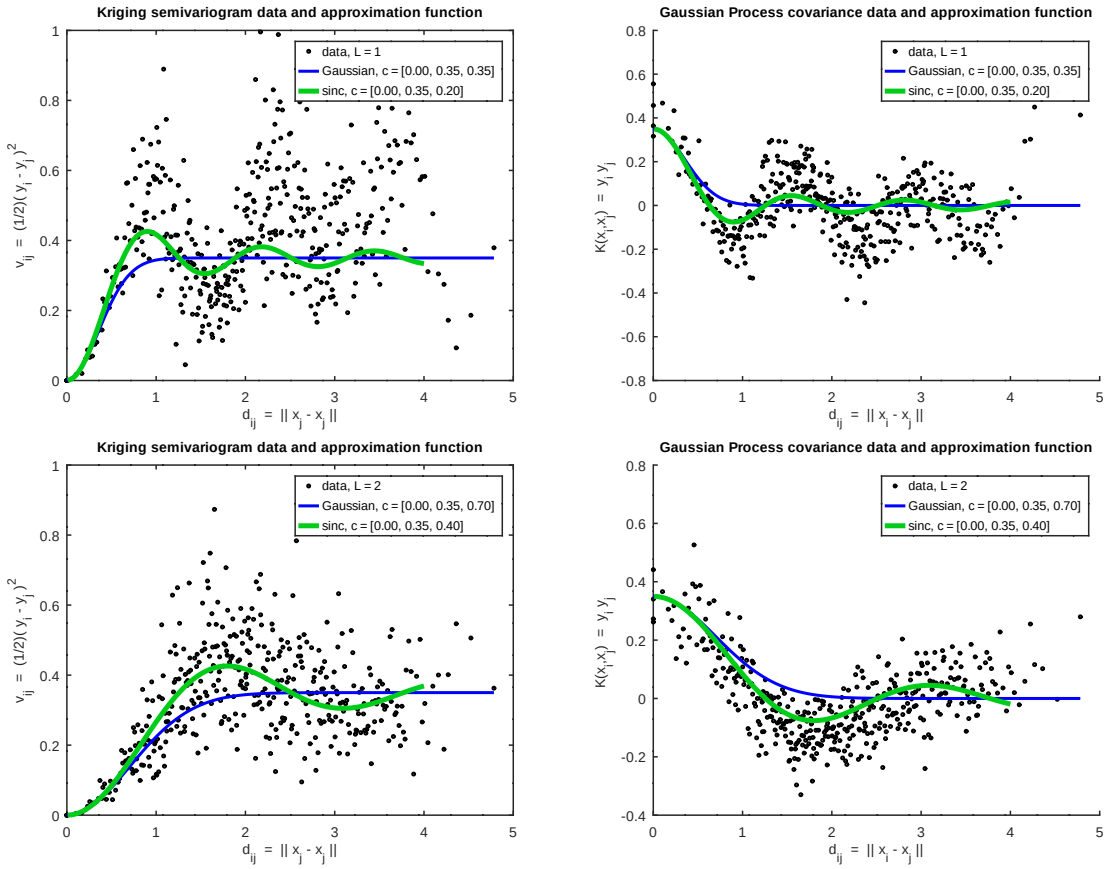

Figure 5. Semivariogram and covariance data computed from 105 randomly-distributed samples in the domain $(-2 < x_1, x_2 < 2)$, and semivariogram functions and covariance functions representative of the data. (a) and (b): length scale $L = 1$, (c) and (d): length scale $L = 2$. Note here that the sinc functions capture the oscillatory trend in the data, and therefore might be expected to provide a more accurate interpolation.

5.2 Results

The surfaces of equation (51), plotted in Figure 4, were interpolated with IDW interpolation with $\alpha = 0.1$ and $q = 3$. The Kriging and Gaussian Process interpolations use sinc and Gaussian semivariogram and covariance functions, as shown in Figure 5. Sample points were either regularly distributed in a triangular mesh pattern or were uniformly distributed over the domain. The interpolation's root mean square error $E_{\rm rms}$ and its correlation with the noise-free function were calculated for length scales of $L = 1$ and $L = 2$, and using $m = 18$, $m = 53$, and $m = 105$ measurement points.

The set of $m = 105$ regularly spaced interpolation points are shown as blue points in Figure 4. Examples of the interpolated surfaces are shown in Figures 6 and 7 for $L = 1$ and $L = 2$, respectively.

A visual comparison of Figures 4, 6, and 7 illustrates, as an example, the relative characteristics of IDW, Kriging, and Gaussian Process interpolation. To further evaluate

the relative characteristics of IDW, Kriging, and Gaussian Process interpolation, the R.M.S. error, equation (52) and the correlation between the interpolated surfaces and the noise-free surface are tabulated for other cases. In the following tables, the symbol "ms" indicates that in the corresponding case the matrix G or K are nearly singular. In these examples the sinc approximation functions resulted in singular matrices for both Kriging and Gaussian Process interpolation. The Gaussian approximation function resulted in well conditioned matrices, except for some of the cases with significant modeled observation noise.

Table 1. gridded observations no noise, no nugget, Sinc semivariogram and covariance approximation function

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.484	0.515	0.764	0.764	0.558	0.541	0.310	0.310
$L = 2$	0.345	0.263	0.311	0.311	0.867	0.880	0.830	0.830
$m = 53$								
$L = 1$	0.484	0.515	0.764	0.764	0.558	0.541	0.310	0.310
$L = 2$	0.508	0.520	ms	ms	0.349	0.273	ms	ms
$m = 105$								
$L = 1$	0.504	0.504	ms	ms	0.264	0.240	ms	ms
$L = 2$	0.169	0.066	ms	ms	0.989	0.991	ms	ms

Table 2. gridded observations no noise, no nugget, Gaussian semivariogram and covariance approximation function Figures 6 and 7

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.484	0.515	0.459	0.459	0.558	0.541	0.554	0.554
$L = 2$	0.526	0.554	0.533	0.533	0.390	0.361	0.320	0.320
$m = 53$								
$L = 1$	0.504	0.504	0.501	0.501	0.264	0.240	0.224	0.224
$L = 2$	0.262	0.236	0.128	0.128	0.877	0.878	0.969	0.969
$m = 105$								
$L = 1$	0.332	0.249	0.127	0.127	0.861	0.864	0.972	0.972
$L = 2$	0.169	0.066	0.001	0.001	0.989	0.991	1.000	1.000

For interpolating sparse noise-free measurements with Kriging and Gaussian Process interpolation the use of uniformly randomly distributed samples appears to provide an interpolation that is closer to the baseline function, as compared to interpolation with samples drawn from a triangular grid. In contrast, the accuracy of IDW interpolation does not appear to be sensitive to the randomness of the distribution of the measurement points.

For interpolating noise-free measurements with Kriging and Gaussian Process interpolation three to five regularly spaced measurement points distributed along the characteristic

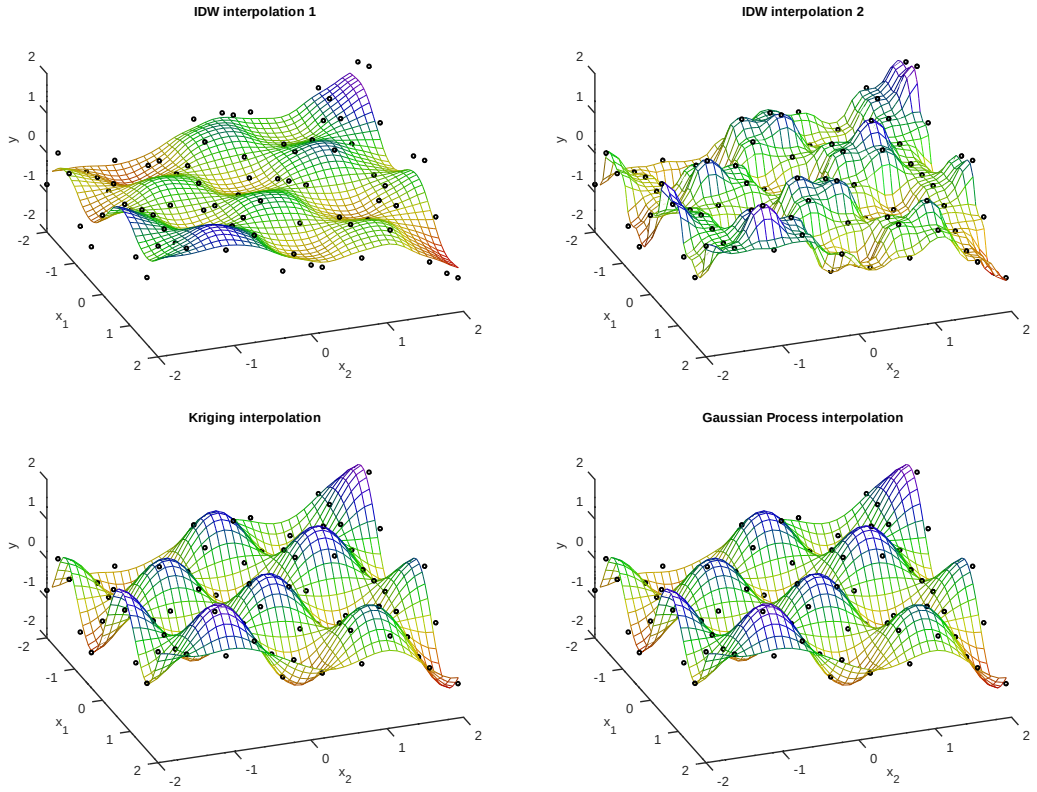

Figure 6. Interpolated surfaces for $L = 1$ computed from $m = 105$ noise-free observations measured at triangular grid points within the domain (Table 2)

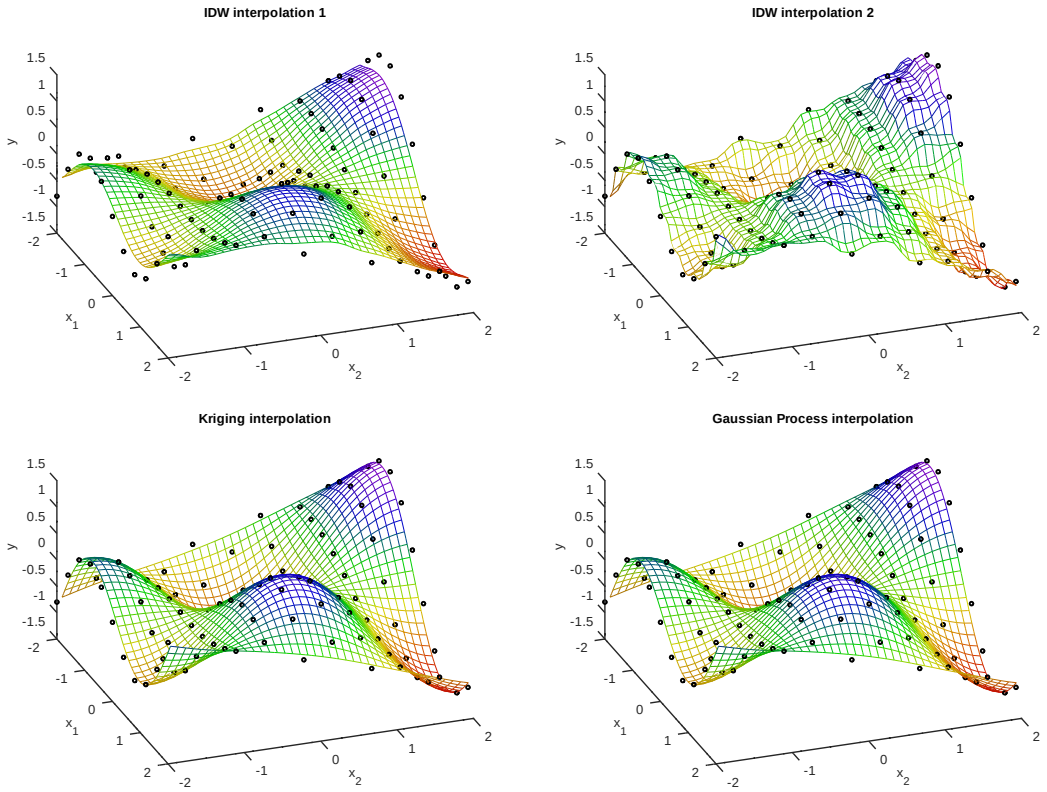

Figure 7. Interpolated surfaces for $L = 2$ computed from $m = 105$ noise-free observations measured at triangular grid points within the domain (Table 2)

Table 3. uniformly distributed random observations, no noise, no nugget, Sinc semivariogram and covariance approximation function

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.434	0.452	ms	ms	0.482	0.456	ms	ms
$L = 2$	0.218	0.219	0.507	0.489	0.892	0.878	0.703	0.713
$m = 53$								
$L = 1$	0.387	0.339	ms	ms	0.637	0.680	ms	ms
$L = 2$	0.325	0.278	ms	ms	0.770	0.813	ms	ms
$m = 105$								
$L = 1$	0.300	0.179	ms	ms	0.853	0.903	ms	ms
$L = 2$	0.205	0.118	ms	ms	0.955	0.969	ms	ms

Table 4. uniformly distributed random observations, no noise, no nugget, Gaussian semivariogram and covariance approximation function, Figures 8 and 9

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.434	0.452	0.366	0.367	0.482	0.456	0.645	0.649
$L = 2$	0.218	0.219	0.173	0.167	0.892	0.878	0.945	0.949
$m = 53$								
$L = 1$	0.387	0.339	0.254	0.253	0.637	0.680	0.824	0.830
$L = 2$	0.325	0.278	0.136	0.125	0.770	0.813	0.973	0.978
$m = 105$								
$L = 1$	0.300	0.179	0.104	0.102	0.853	0.903	0.979	0.980
$L = 2$	0.205	0.118	0.012	0.012	0.955	0.969	1.000	1.000

Table 5. uniformly distributed random observations, noise: $\sigma_N = 0.5$, no nugget, Sinc semivariogram and covariance approximation function

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.450	0.443	ms	ms	0.489	0.482	ms	ms
$L = 2$	0.342	0.414	1.017	1.040	0.767	0.754	0.415	0.426
$m = 53$								
$L = 1$	0.386	0.335	ms	ms	0.507	0.558	ms	ms
$L = 2$	0.371	0.332	ms	ms	0.690	0.693	ms	ms
$m = 105$								
$L = 1$	0.300	0.233	ms	ms	0.730	0.762	ms	ms
$L = 2$	0.231	0.199	ms	ms	0.867	0.829	ms	ms

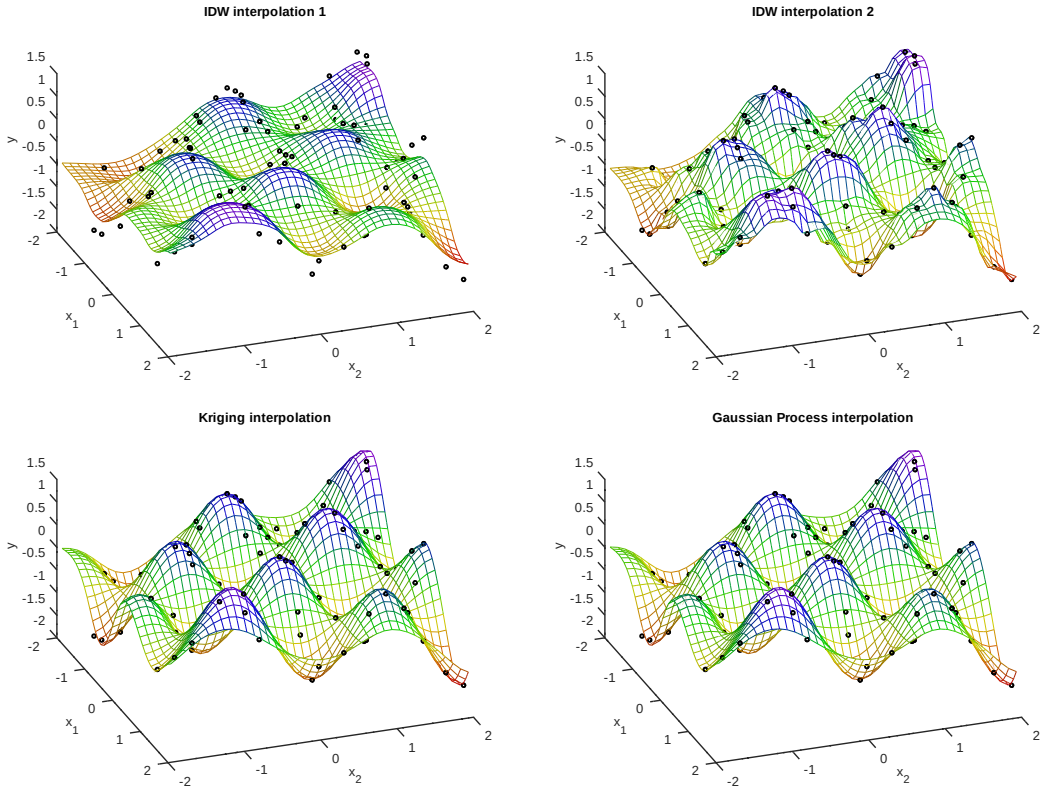

Figure 8. Interpolated surfaces for $L = 1$ computed from $m = 105$ noise-free observations measured at random points uniformly distributed over the domain (Table 4)

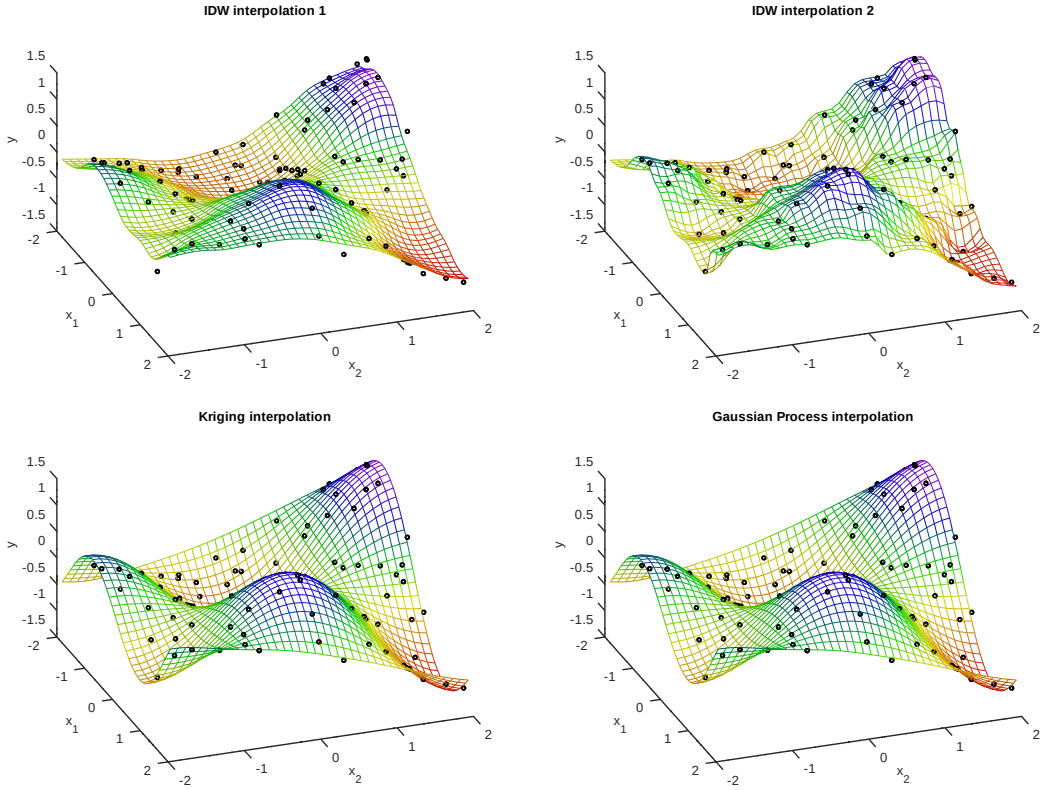

Figure 9. Interpolated surfaces for $L = 2$ computed from $m = 105$ noise-free observations measured at random points uniformly distributed over the domain (Table 4)

Table 6. uniformly distributed random observations, noise: $\sigma_N = 0.5$, no nugget, Gaussian semivariogram and covariance approximation function

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.450	0.443	0.405	0.405	0.489	0.482	0.589	0.591
$L = 2$	0.342	0.414	0.398	0.394	0.767	0.754	0.802	0.805
$m = 53$								
$L = 1$	0.386	0.335	0.303	0.307	0.507	0.558	0.641	0.645
$L = 2$	0.371	0.332	0.781	0.795	0.690	0.693	0.331	0.349
$m = 105$								
$L = 1$	0.300	0.233	0.878	0.876	0.730	0.762	0.300	0.304
$L = 2$	0.231	0.199	ms	ms	0.867	0.829	ms	ms

Table 7. uniformly distributed random observations, noise: $\sigma_N = 0.5$, nugget: $c_1 = 0.10$, Sinc semivariogram and covariance approximation function

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.450	0.443	ms	0.512	0.489	0.482	ms	0.296
$L = 2$	0.342	0.414	ms	0.494	0.767	0.754	ms	0.445
$m = 53$								
$L = 1$	0.386	0.335	ms	0.496	0.507	0.558	ms	0.249
$L = 2$	0.371	0.332	ms	0.472	0.690	0.693	ms	0.468
$m = 105$								
$L = 1$	0.300	0.233	ms	0.477	0.730	0.762	ms	0.355
$L = 2$	0.231	0.199	ms	0.340	0.867	0.829	ms	0.755

Table 8. uniformly distributed random observations, noise: $\sigma_N = 0.5$, nugget: $c_1 = 0.10$, Gaussian semivariogram and covariance approximation function, Figures 10 and 11

	RMS Error				Correlation Coefficients			
	IDW1	IDW2	Kriging	G.P.	IDW1	IDW2	Kriging	G.P.
$m = 18$								
$L = 1$	0.450	0.443	0.405	0.425	0.489	0.482	0.589	0.583
$L = 2$	0.342	0.414	0.398	0.320	0.767	0.754	0.802	0.783
$m = 53$								
$L = 1$	0.386	0.335	0.303	0.313	0.507	0.558	0.641	0.626
$L = 2$	0.371	0.332	0.781	0.265	0.690	0.693	0.331	0.825
$m = 105$								
$L = 1$	0.300	0.233	0.878	0.252	0.730	0.762	0.300	0.803
$L = 2$	0.231	0.199	ms	0.224	0.867	0.829	ms	0.877

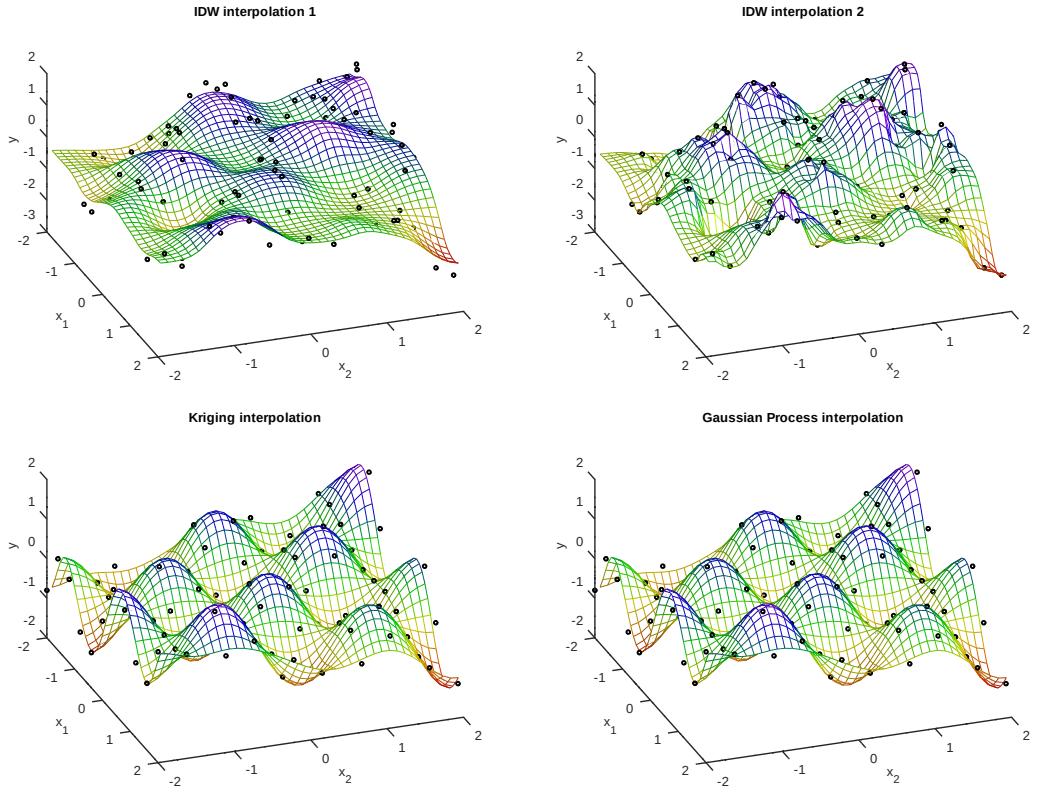

Figure 10. Interpolated surfaces for $L = 1$ computed from $m = 105$ noisy observations measured at random points uniformly distributed over the domain (Table 8)

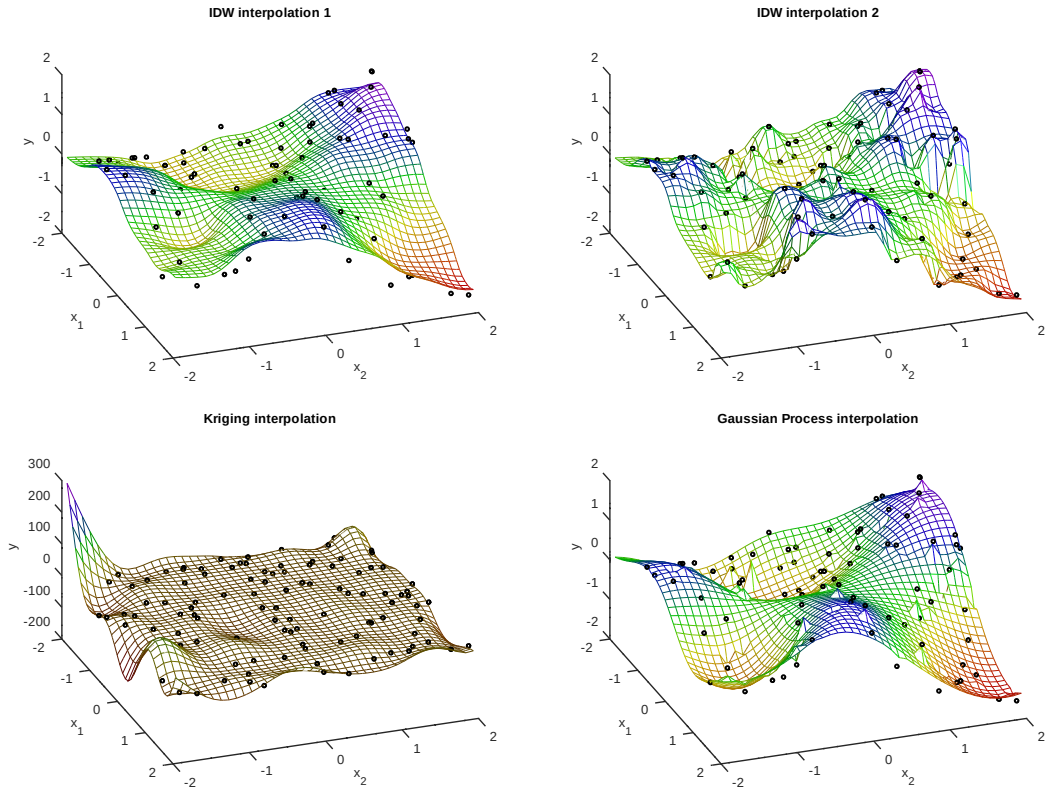

Figure 11. Interpolated surfaces for $L = 2$ computed from $m = 105$ noisy observations measured at random points uniformly distributed over the domain (Table 8)

length scale of the features of the surface is sufficient for Kriging interpolation to be accurate to within one percent. The accuracy of IDW interpolation with randomly or regularly spaced measurements is roughly a tenth of the accuracy of Kriging with regularly spaced measurements.

The distribution of interpolation errors is easily estimated in Kriging and Gaussian Process interpolation and enables an identification of regions of high accuracy and regions of low accuracy. Regions of high accuracy naturally coincide with regions of denser measurement points.

6 Anisotropic semivariogram and kernel functions

The three interpolation methods described here may be generalized to interpolate data with anisotropic variability by simply replacing the Euclidian distance (squared, and scaled by c_3)

$$d_{ij}^2/c_3^2 = (x_i - x_j)^\mathsf{T}(x_i - x_j) \;/\; c_3^2$$

with a squared distance variable that is weighted by the inverse of the covariance matrix $V \in \mathbb{R}^{n\times n}$ ($V = V^\mathsf{T}, V > 0$) of the anisotropic variability,

$$d_{ij}^2/c_3^2 = (x_i - x_j)^\mathsf{T} V^{-1}(x_i - x_j).$$

For convienence, we define the difference vector between two sets of features

$$p_{k(i,j)} = x_i - x_j$$

with x_i, x_j, and $p_k \in \mathbb{R}^n$, $i, j \in \{1, \cdots, m\}$. Defining the matrix V as the covariance matrix of a Gaussian kernel function with a mean function of zero,

$$\hat\kappa(x_i, x_j; V^{-1}) = c_2 \exp\left[\, -\frac{1}{2}(x_i - x_j)^\mathsf{T} V^{-1}(x_i - x_j) \,\right] = c_2 \exp\left[\, -\frac{1}{2}\, p_k^\mathsf{T}\, V^{-1}\, p_k \,\right]$$

The corresponding value of the kernel $\kappa(x_i, x_j)$ is the covariance $\mathbb{E}[Y(x_i)Y(x_j)]$ (for a mean function $m(x)$ of zero). The coefficient c_2 is simply the variance $\mathbb{E}[Y(x_i)Y(x_i)]$. A maximum likelihood estimate for the inverse covariance matrix V^{-1} may be found without iteration using log-transformed variables. The error in the log-transformed distribution is

$$e_k = \ln \hat\kappa(x_i, x_j; V^{-1}) - \ln \kappa(x_i, x_j) = \ln c_2 - \frac{1}{2}\left(p_k^\mathsf{T} V^{-1} p_k\right) - \ln \kappa(x_i, x_j)\ . \tag{53}$$

The error e_k is linear in the coefficients of V^{-1},

$$p_k^\mathsf{T} V^{-1} p_k = \bar p_k^\mathsf{T} \bar v \tag{54}$$

where, for symmetric covariance matrices, $\bar v$ and $\bar p_k$ are vectors of length $n \times (n+1)/2$.

$$\bar v = \begin{bmatrix} [V_{11}^{-1} & V_{12}^{-1} & \cdots & V_{1n}^{-1}]^\mathsf{T} \\ [V_{22}^{-1} & V_{23}^{-1} & \cdots & V_{2n}^{-1}]^\mathsf{T} \\ [V_{33}^{-1} & V_{34}^{-1} & \cdots & V_{3n}^{-1}]^\mathsf{T} \\ \vdots & & & \\ V_{nn}^{-1} & & & \end{bmatrix}_{(n(n+1)/2)\times 1}$$

and

$$\bar{p}_k = \begin{bmatrix} p_{k1} & [p_{k1} & 2p_{k2} & \cdots & 2p_{kn}]^\mathsf{T} \\ p_{k2} & [p_{k2} & 2p_{k3} & \cdots & 2p_{kn}]^\mathsf{T} \\ p_{k3} & [p_{k3} & 2p_{k4} & \cdots & 2p_{kn}]^\mathsf{T} \\ \vdots & & & \ddots & \\ p_{kn} & [& & & p_{kn}] \end{bmatrix}_{(n(n+1)/2)\times 1}$$

Defining $\bar{P} \in \mathbb{R}^{(n(n+1)/2)\times m^2}$ as the concatination of all m^2 vectors $\bar{p}_k$

$$\bar{P} = \begin{bmatrix} \bar{p}_1 & \bar{p}_2 & \bar{p}_3 & \cdots & \bar{p}_{m^2} \end{bmatrix} ,$$

the set of m^2 errors $\bar{e} = [e_1, e_2, \cdots, e_{m^2}]^\mathsf{T}$ becomes

$$\bar{e} = \ln c_2 - \frac{1}{2}\bar{P}^\mathsf{T}\bar{v} - \ln \kappa$$

from which, under the assumption of normally distributed errors, the maximum likelihood estimate of $\bar{v}$ is the orthogonal projection of $2(\ln c_2 - \ln \kappa)$ onto the basis comprised by rows of $\bar{P}$, and solves

$$[\bar{P}\bar{P}^\mathsf{T}]\ \bar{v} = P\ 2(\ln c_2 - \ln \kappa)$$

In practice, values of κ are computed from estimates of $\mathbb{E}[Y_i Y_j]$ by sampling within each MECE hyperbox in the n-dimensional parameter space. When a sample of a Gaussian Process in which the domain of the independent variables is not much larger than the largest eigenvalue of V^{-1}, values of $\kappa(x_i, x_j)$ may be negative, and can not be log-transfored. In such cases, which are not uncommon, a subset of the m^2 columns of $\bar{P}$ are chosen as the cluster of $p(x_i, x_j)$ values centered at the origin and for which $\kappa(x_i, x_j) > \epsilon > 0$. This is a simple application of distribution-based clustering.

7 Examples

8 Summary

Sparsely-sampled measurements of several independent variables may be interpolated by calculating a weighted average of the measurements. In inverse distance weighted (IDW) interpolation the weights decrease monotonically with the distance between the interpolation point and the measurement point. In least-squares-optimal interpolation (Kriging), the weights minimize a squared error criterion which can be expressed as a function of the squared difference between measured values, the so-called semivariogram. By approximating the semivariogram as an analytic function of the distance separating measurement points, the Kriging weights can be evaluated in closed-form. The interpolation accuracy is readily computed and provides a means of assessing the distribution of interpolated values.

In this study IDW, Kriging, and Gaussian process interpolation methods are applied to two surfaces with differing characteristic length scales.

Even when observation noise is significant ($SNR \approx 2$) Gaussian process interpolation using a nugget in the covariance approximation function and using at least five sample points per the characteristic length scale can recover the noise-free relationship with roughly 85 percent correlation.

9 References

1. Ziko Coulter and Chuong Do, Linear Algebra Review and Reference,
CS 229 Machine Learning, Stanford University,
http://cs229.stanford.edu/section/cs229-linalg.pdf, September 30, 2015

2. Arian Maleki and Tom Do, Review of Probability,
CS 229 Machine Learning, Stanford University,
http://cs229.stanford.edu/section/cs229-prob.pdf

3. Taide Ding, Fereshte Khani, Probability Theory Review,
CS 229 Machine Learning, Stanford University,
http://cs229.stanford.edu/section/cs229_spring2020_prob_review_slides.pdf,April 17, 2020

4. Chuong B. Do, The Multivariate Gaussian Distribution,
CS 229 Machine Learning, Stanford University,
http://cs229.stanford.edu/section/cs229-gaussians.pdf, October 10, 2008

5. Chuong B. Do, More on Multivariate Gaussians,
CS 229 Machine Learning, Stanford University,
http://cs229.stanford.edu/section/cs229-moregaussians.pdf, November 21, 2008

6. Chuong B. Do and Honglak Lee, Gaussian Processes,
CS 229 Machine Learning, Stanford University,
http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf, November 22, 2008

7. Cressie, Noel A. C., *Statistics for Spatial Data,* Wiley, 1991.

8. Deutsch, C. V. and Journel, A. G (1992). GSLIB Geostatistical Software Library and User's Guide. New York, NY: Oxford University Press, Inc., pp. 61-115.

9. Davis, J. C. and McCullagh, M. J. (1975). *Display and Analysis of Spatial Data.* Bristol, Great Britain: J. W. Arrowsmith Ltd., pp. 96-114.

10. Davis, J. C. (1986). *Statistics and Data Analysis in Geology,*

11. Lophaven, S., Nielsen, H., and , Sondergaard, J., *DACE: A , Matlab Kriging Toolbox, Version 2.0.* (2002), Informatics and Mathematical Modeling, Technical University of Denmark.
http://www.imm.dtu.dk/~hbn/dacewww.dace

12. Matherson, G., "Principles of Geostatistics," *Economic Geology,* 58: 1246-1266 (1963)

13. Carl Edward Rasmussen and Christopher K. I. Williams *Gaussian Processes for Machine Learning,* The MIT Press, 2006. ISBN 0-262-18253-X. http://www.gaussianprocess.org/gpml/

14. Swan, A.R.H. and Sandilands, M. (1995) *Introduction to Geological Data Analysis,* Oxford: Blackwell Science, Ltd., 446 pages. Second Edition. John Wiley and Sons, pp. 383-403. QE48.8 .D38 1986