Some Recent Insights into Computing with Positive Definite Kernels

Greg Fasshauer

Department of Applied Mathematics & Statistics
Colorado School of Mines
Partially supported by NSF Grant DMS–1522687

Joint work with Mike McCourt

MAIA Luminy, September 21, 2016

Outline

Outline

New Perspectives

Illinois

New Perspectives

Colorado

Deterministic Kernel-based Interpolation

Given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, use a data-dependent linear function space, i.e.,

$$s(\boldsymbol{x}) = \sum_{j=1}^{N} c_j K(\boldsymbol{x}, \boldsymbol{x}_j) = \boldsymbol{k}(\boldsymbol{x})^T \boldsymbol{c}, \qquad \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^d$$

with $K : \Omega \times \Omega \to \mathbb{R}$ a positive definite reproducing kernel.

Deterministic Kernel-based Interpolation

Given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, use a data-dependent linear function space, i.e.,

$$s(\boldsymbol{x}) = \sum_{j=1}^{N} c_j K(\boldsymbol{x}, \boldsymbol{x}_j) = \boldsymbol{k}(\boldsymbol{x})^T \boldsymbol{c}, \qquad \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^d$$

with $K : \Omega \times \Omega \to \mathbb{R}$ a positive definite reproducing kernel.
To find c_j solve the interpolation equations

$$s(\boldsymbol{x}_i) = y_i, \quad i = 1, \ldots, N,$$

which leads to a linear system $\mathrm{K}\boldsymbol{c} = \boldsymbol{y}$ with symmetric positive definite — often ill-conditioned — system matrix

$$\mathrm{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad i, j = 1, \ldots, N,$$

Deterministic Kernel-based Interpolation

Given data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, use a data-dependent linear function space, i.e.,

$$s(\boldsymbol{x}) = \sum_{j=1}^{N} c_j K(\boldsymbol{x}, \boldsymbol{x}_j) = \boldsymbol{k}(\boldsymbol{x})^T \boldsymbol{c}, \qquad \boldsymbol{x} \in \Omega \subseteq \mathbb{R}^d$$

with $K : \Omega \times \Omega \to \mathbb{R}$ a positive definite reproducing kernel.
To find c_j solve the interpolation equations

$$s(\boldsymbol{x}_i) = y_i, \quad i = 1, \ldots, N,$$

which leads to a linear system $\mathrm{K}\boldsymbol{c} = \boldsymbol{y}$ with symmetric positive definite — often ill-conditioned — system matrix

$$\mathrm{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad i, j = 1, \ldots, N,$$

and cardinal representation

$$s(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1} \boldsymbol{y} = \ell(\boldsymbol{x})^T \boldsymbol{y}.$$

Connection to Kriging

Given a Gaussian (zero-mean) random field Y with covariance kernel K, process variance σ^2 and observations

$$\boldsymbol{Y} = \begin{pmatrix} Y_{\boldsymbol{x}_1} & \cdots & Y_{\boldsymbol{x}_N} \end{pmatrix}^T, \quad Y_{\boldsymbol{x}_j} \text{ zero-mean random variables},$$

the (simple) kriging predictor is of the form

$$\widehat{Y}_{\boldsymbol{x}} = \sum_{j=1}^{N} w_j(\boldsymbol{x}) Y_{\boldsymbol{x}_j} = \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y},$$

- $\widehat{Y}_{\boldsymbol{x}}$: zero-mean random variable,
- $\boldsymbol{w}(\cdot) = \begin{pmatrix} w_1(\cdot) & \cdots & w_N(\cdot) \end{pmatrix}^T$: vector of weight functions.

"Optimal" weights $\overset{\star}{w}_j(\cdot)$ will minimize the MSE of the predictor, i.e.,

$$\text{MSE}(\widehat{Y}_{\boldsymbol{x}}) = \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\right)^2\right].$$

Using the covariance kernel K and process variance σ^2 of Y, i.e., $\sigma^2 K(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}]$, we have

$$\begin{aligned}
\text{MSE}(\widehat{Y}_{\boldsymbol{x}}) &= \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\right)^2\right] \\
&= \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{x}}] - 2\mathbb{E}[Y_{\boldsymbol{x}} \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}] + \mathbb{E}[\boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}(\boldsymbol{x})] \\
&= \sigma^2 K(\boldsymbol{x}, \boldsymbol{x}) - 2\boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \boldsymbol{k}(\boldsymbol{x})) + \boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \mathrm{K})\boldsymbol{w}(\boldsymbol{x}).
\end{aligned}$$

Using the covariance kernel K and process variance σ^2 of Y, i.e., $\sigma^2 K(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}]$, we have

$$\begin{aligned}
\text{MSE}(\widehat{Y}_{\boldsymbol{x}}) &= \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\right)^2\right] \\
&= \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{x}}] - 2\mathbb{E}[Y_{\boldsymbol{x}} \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}] + \mathbb{E}[\boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}(\boldsymbol{x})] \\
&= \sigma^2 K(\boldsymbol{x}, \boldsymbol{x}) - 2\boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \boldsymbol{k}(\boldsymbol{x})) + \boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \mathrm{K})\boldsymbol{w}(\boldsymbol{x}).
\end{aligned}$$

Differentiation and equating to 0 yields the optimum weight vector

$$\overset{\star}{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}),$$

Using the covariance kernel K and process variance σ^2 of Y, i.e., $\sigma^2 K(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{z}}]$, we have

$$\begin{aligned}\text{MSE}(\widehat{Y}_{\boldsymbol{x}}) &= \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\right)^2\right] \\ &= \mathbb{E}[Y_{\boldsymbol{x}} Y_{\boldsymbol{x}}] - 2\mathbb{E}[Y_{\boldsymbol{x}} \boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}] + \mathbb{E}[\boldsymbol{w}(\boldsymbol{x})^T \boldsymbol{Y}\boldsymbol{Y}^T \boldsymbol{w}(\boldsymbol{x})] \\ &= \sigma^2 K(\boldsymbol{x}, \boldsymbol{x}) - 2\boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \boldsymbol{k}(\boldsymbol{x})) + \boldsymbol{w}(\boldsymbol{x})^T(\sigma^2 \mathrm{K})\boldsymbol{w}(\boldsymbol{x}).\end{aligned}$$

Differentiation and equating to 0 yields the optimum weight vector

$$\overset{\star}{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}),$$

so that the (simple) kriging predictor

$$\widehat{Y}_{\boldsymbol{x}} = \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1} \boldsymbol{Y}$$

is the best linear unbiased predictor.

Remark

Note that this is independent of the process variance σ^2.

Parametrized Kernel Methods

Many kernels contain parameters whose values greatly affect the performance of the kernel method.

For example, such parameters may affect

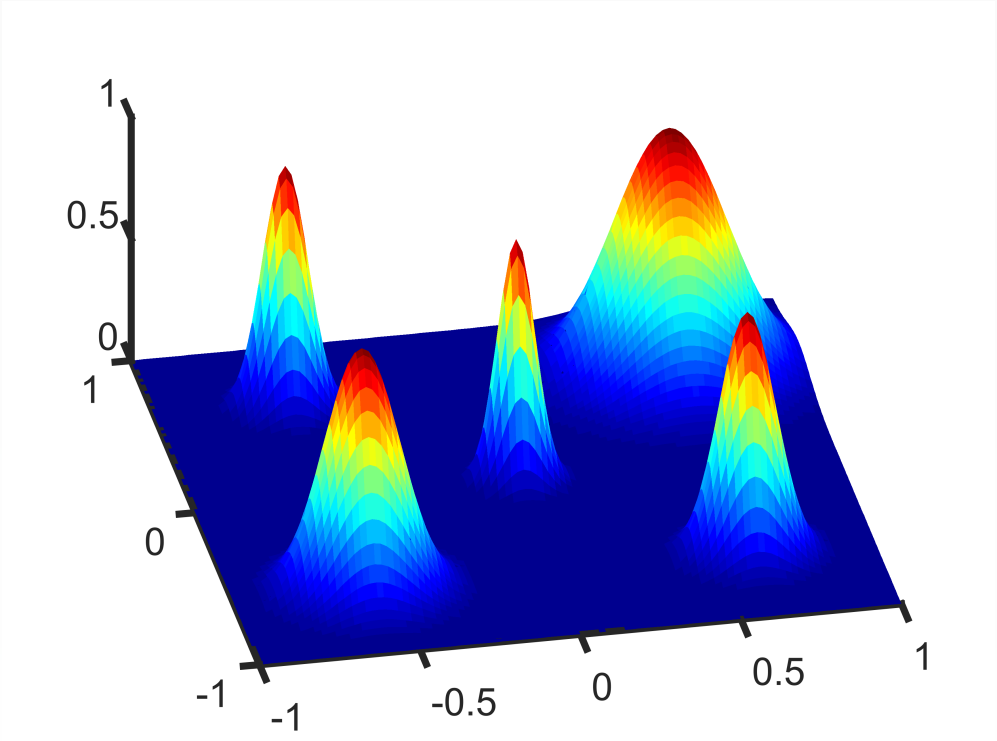
- shape,
- smoothness,
- accuracy,
- numerical stability,
- computational efficiency, i.e., density of (sparse) K,
- balance between closeness of fit and smoothness.

It is therefore important to find "optimal" values for these parameters.

Examples

- isotropic shape parameter, e.g.,

$$K(\boldsymbol{x}, \boldsymbol{z}) = \kappa(r) = \mathrm{e}^{-\varepsilon^2 r^2}, \quad r = \|\boldsymbol{x} - \boldsymbol{z}\|_2$$

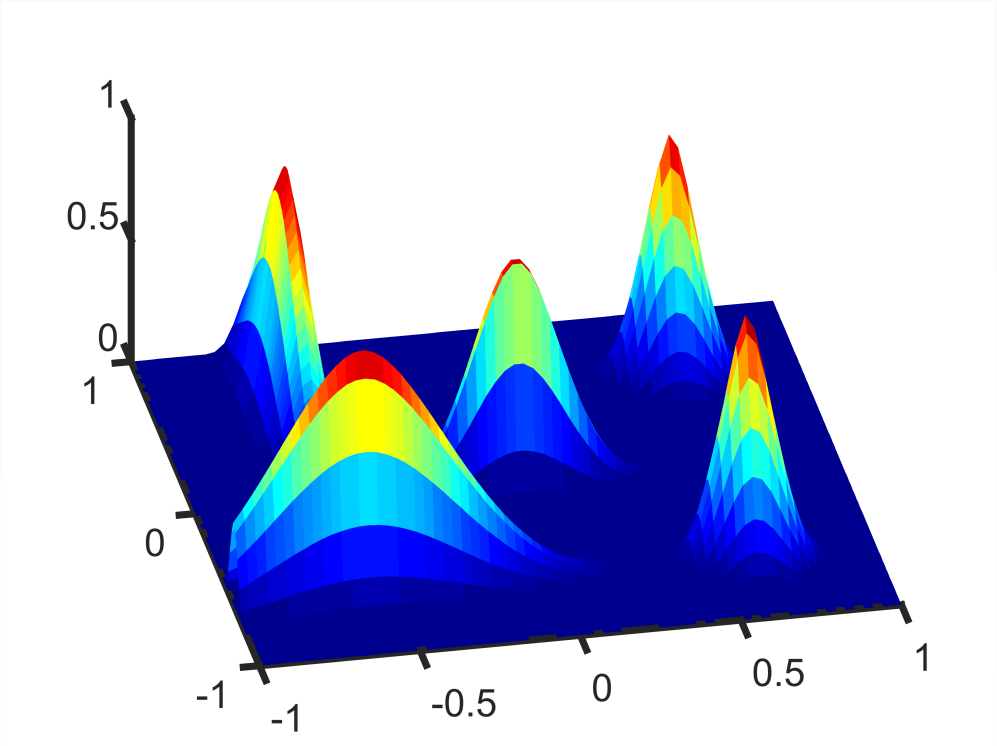


Isotropic Gaussian kernels, $\varepsilon = 4, 6, 8, 10, 12$

Examples (cont.)

- radial anisotropic shape parameters, e.g.,

$$\kappa(r) = (1-r)_+^4(4r+1), \quad r = \sqrt{(\boldsymbol{x}-\boldsymbol{z})^T \mathrm{E}(\boldsymbol{x}-\boldsymbol{z})}, \quad \mathrm{E} = \begin{pmatrix} \varepsilon_1^2 & & \\ & \ddots & \\ & & \varepsilon_d^2 \end{pmatrix}$$

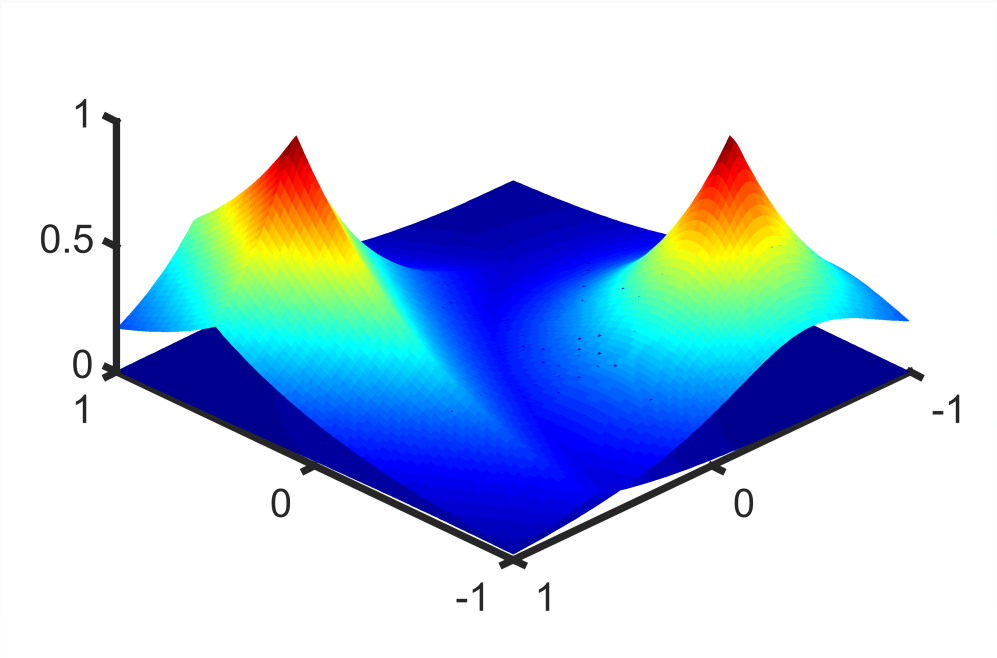


Compactly supported Wendland kernels,
$\varepsilon = [4,4], [2,2], [5,1], [1,5], [2,7]$

Examples (cont.)

- tensor product anisotropic shape parameters, e.g.,

$$K(\boldsymbol{x}, \boldsymbol{z}) = \prod_{\ell=1}^{d} \mathrm{e}^{-\varepsilon_\ell |x_\ell - z_\ell|}$$



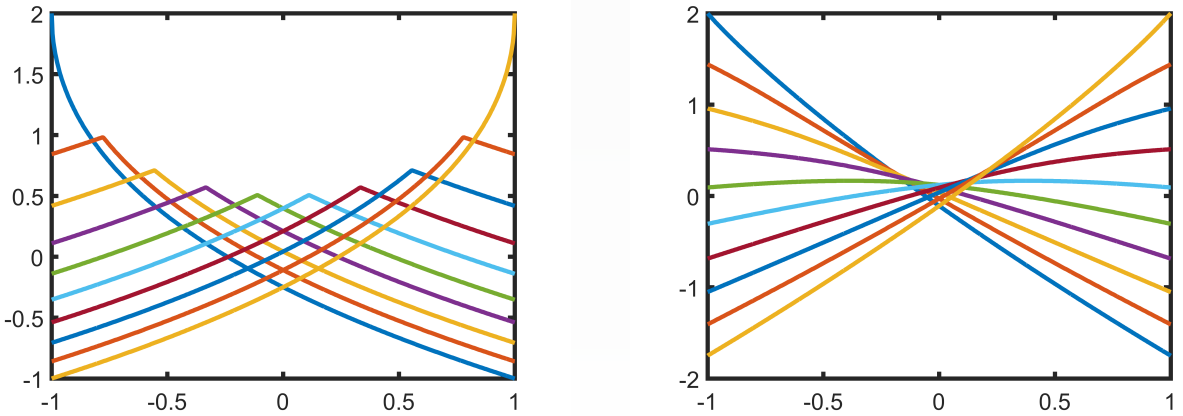
Tensor product and radial C^0 Matérn kernels,
$\varepsilon = [1.5, 2.5], \varepsilon = 2.5$

Examples (cont.)

- smoothness parameter(s), e.g.,

$$K(x,z) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(x)\varphi_n(z)$$

$$\lambda_n = \frac{1}{\zeta(2\beta)n^{2\beta}}, \quad \varphi_n(x) = \sqrt{2}T_n(x), \quad n = 1,2,\ldots$$



Chebyshev kernels with $\beta = 1, 2$

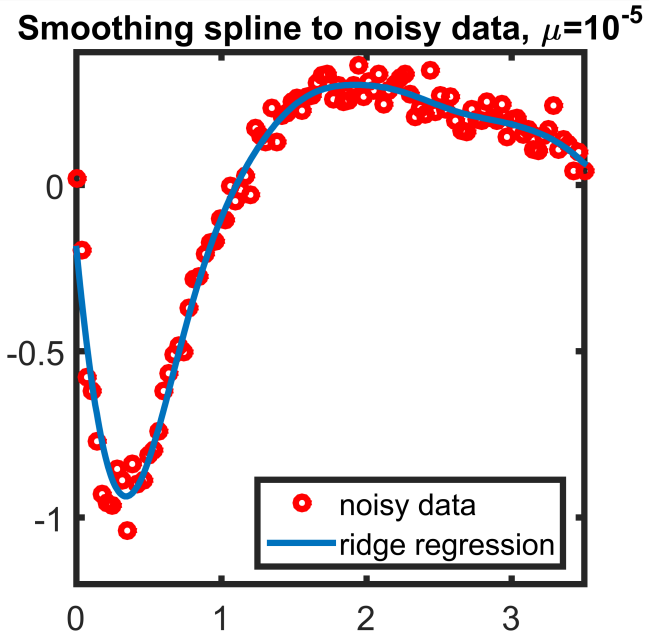
Examples (cont.)

- regularization parameter, e.g.,

$$\min_{\boldsymbol{c}\in\mathbb{R}^N} \left[(\boldsymbol{y} - \mathrm{K}\boldsymbol{c})^T (\boldsymbol{y} - \mathrm{K}\boldsymbol{c}) + \mu \boldsymbol{c}^T \mathrm{K} \boldsymbol{c} \right]$$

so that we have the smoothing spline/ridge regression

$$s(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^T (\mathrm{K} + \mu \mathrm{I})^{-1} \boldsymbol{y}$$



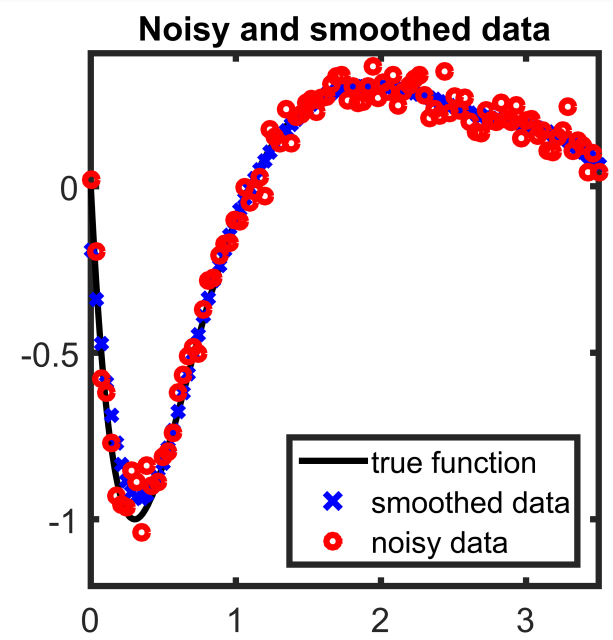
Examples (cont.)

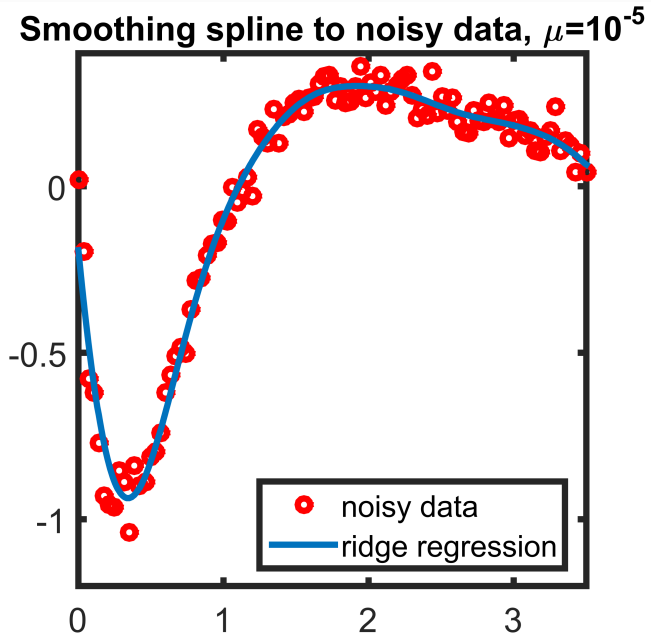
- regularization parameter, e.g.,

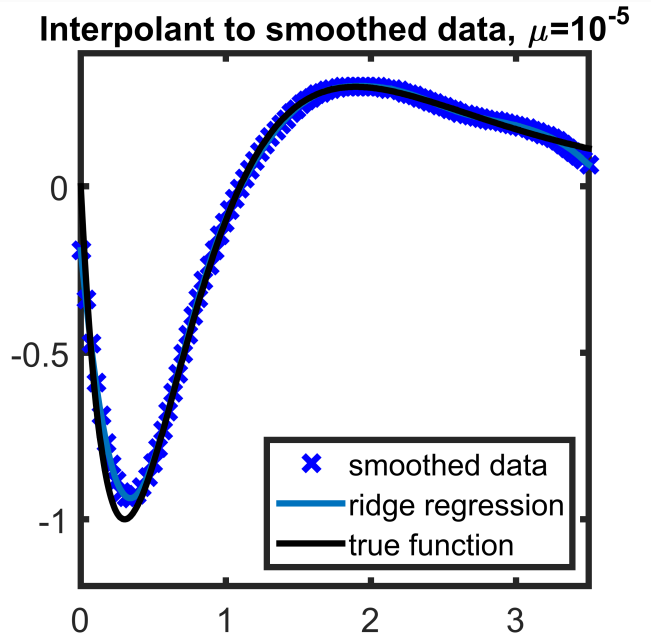
$$\min_{\boldsymbol{c}\in\mathbb{R}^N} \left[(\boldsymbol{y} - \mathrm{K}\boldsymbol{c})^T (\boldsymbol{y} - \mathrm{K}\boldsymbol{c}) + \mu \boldsymbol{c}^T \mathrm{K} \boldsymbol{c} \right]$$

so that we have the smoothing spline/ridge regression

$$\begin{aligned} s(\boldsymbol{x}) &= \boldsymbol{k}(\boldsymbol{x})^T (\mathrm{K} + \mu \mathrm{I})^{-1} \boldsymbol{y} = \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1} (\mathrm{K} + \mu \mathrm{I})^{-1} \mathrm{K} \boldsymbol{y} \\ &= \ell(\boldsymbol{x})^T \tilde{\boldsymbol{y}}. \end{aligned}$$







Outline

How do we decide "optimality"?

We use parametrization criteria such as

$$\mathrm{C_{CV}}(\varepsilon;p) = \left\| \left(\frac{\boldsymbol{c}_1}{\mathrm{K}_{11}^{-1}} \quad \cdots \quad \frac{\boldsymbol{c}_N}{\mathrm{K}_{NN}^{-1}} \right) \right\|_p, \qquad \text{LOOCV}$$

$$\mathrm{C_{MPLE}}(\varepsilon) = N \log \left(\boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y} \right) + \log\det \mathrm{K}, \qquad \text{profile likelihood}$$

$$\mathrm{C_{GW}}(\varepsilon;p) = \sqrt{\boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}}\, \|P_{K,\mathcal{X}}\|_p, \qquad \text{Golomb–Weinberger}$$

where $P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1} \boldsymbol{k}(\boldsymbol{x})}$ or kriging variance

These criteria all aim to maximize some form of accuracy.
They may require computing

$$\mathrm{K}^{-1}\boldsymbol{y}, \quad \boldsymbol{y}^T\mathrm{K}^{-1}\boldsymbol{y}, \quad \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}, \quad \log\det \mathrm{K}, \quad P_{K,\mathcal{X}}(\boldsymbol{x}).$$

Outline for remainder of talk

- Explain parametrization criteria:
  - maximum profile likelihood
  - Golomb–Weinberger error bound (kriging variance)
- Explain how to compute
  - cardinal functions $\ell(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T \mathsf{K}^{-1}$
  - log det K
  - power function $P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T \mathsf{K}^{-1} \boldsymbol{k}(\boldsymbol{x})}$
- Show a few numerical examples

Maximum Profile Likelihood

We consider a zero-mean Gaussian random field Y with covariance kernel $\sigma^2 K$ and likelihood function

$$L(\sigma^2, \varepsilon) = \frac{1}{\sqrt{(2\pi)^N \sigma^{2N} \det(\mathrm{K})}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}\right),$$

where ε appears within K.

Maximum Profile Likelihood

We consider a zero-mean Gaussian random field Y with covariance kernel $\sigma^2 K$ and likelihood function

$$L(\sigma^2, \varepsilon) = \frac{1}{\sqrt{(2\pi)^N \sigma^{2N} \det(\mathrm{K})}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}\right),$$

where ε appears within K.

Maximizing $L(\sigma^2, \varepsilon)$ is equivalent to minimizing

$$-2 \log\left(L(\sigma^2, \varepsilon)\right) = N \log 2\pi + N \log \sigma^2 + \log \det \mathrm{K} + \frac{1}{\sigma^2} \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}.$$

Maximum Profile Likelihood

We consider a zero-mean Gaussian random field Y with covariance kernel $\sigma^2 K$ and likelihood function

$$L(\sigma^2, \varepsilon) = \frac{1}{\sqrt{(2\pi)^N \sigma^{2N} \det(\mathrm{K})}} \exp\left(-\frac{1}{2\sigma^2} \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}\right),$$

where ε appears within K.

Maximizing $L(\sigma^2, \varepsilon)$ is equivalent to minimizing

$$-2 \log\left(L(\sigma^2, \varepsilon)\right) = N \log 2\pi + N \log \sigma^2 + \log \det \mathrm{K} + \frac{1}{\sigma^2} \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}.$$

Differentiating with respect to σ^2 and equating to 0 gives the optimal profile variance

$$\sigma^2_{\text{opt}} = \frac{1}{N} \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}.$$

Using the optimal profile variance $\sigma^2_{\text{opt}} = \frac{1}{N}\boldsymbol{y}^T \mathrm{K}^{-1}\boldsymbol{y}$, we now find ε that minimizes

$$
\begin{aligned}
&-2\log\Big(L(\sigma^2_{\text{opt}},\varepsilon)\Big) - N\log 2\pi \\
&\qquad = N\log\sigma^2_{\text{opt}} + \log\det \mathrm{K} + \frac{1}{\sigma^2_{\text{opt}}}\boldsymbol{y}^T\mathrm{K}^{-1}\boldsymbol{y} \\
&\qquad = N\log\Big(\boldsymbol{y}^T\mathrm{K}^{-1}\boldsymbol{y}\Big) - N\log N + \log\det \mathrm{K} + N.
\end{aligned}
$$

Using the optimal profile variance $\sigma^2_{\text{opt}} = \frac{1}{N}\boldsymbol{y}^T \mathrm{K}^{-1}\boldsymbol{y}$, we now find ε that minimizes

$$\begin{aligned}
&-2\log\left(L(\sigma^2_{\text{opt}}, \varepsilon)\right) - N\log 2\pi \\
&\quad = N\log\sigma^2_{\text{opt}} + \log\det \mathrm{K} + \frac{1}{\sigma^2_{\text{opt}}}\boldsymbol{y}^T\mathrm{K}^{-1}\boldsymbol{y} \\
&\quad = N\log\left(\boldsymbol{y}^T\mathrm{K}^{-1}\boldsymbol{y}\right) - N\log N + \log\det \mathrm{K} + N.
\end{aligned}$$

This yields the profile likelihood criterion

$$\mathrm{C}_{\text{MPLE}}(\varepsilon) = N\log\left(\boldsymbol{y}^T\mathrm{K}^{-1}\boldsymbol{y}\right) + \log\det \mathrm{K}.$$

Golomb–Weinberger/kriging variance criterion

Using the representation $\ell(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1}$ of the cardinal functions and

$$\boldsymbol{y} = \begin{pmatrix} f(\boldsymbol{x}_1) & \cdots & f(\boldsymbol{x}_N) \end{pmatrix}^T$$
$$\overset{K \text{ is RK}}{=} \begin{pmatrix} \langle f, K(\cdot, \boldsymbol{x}_1)\rangle_{\mathcal{H}_K} & \cdots & \langle f, K(\cdot, \boldsymbol{x}_N)\rangle_{\mathcal{H}_K} \end{pmatrix}^T = \langle f, \boldsymbol{k}(\cdot)\rangle_{\mathcal{H}_K}$$

we have the standard pointwise error bound for interpolation

$$\begin{aligned} |f(\boldsymbol{x}) - s(\boldsymbol{x})| &= \left| f(\boldsymbol{x}) - \ell(\boldsymbol{x})^T \boldsymbol{y} \right| = \left| \langle f, K(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}_K} - \ell(\boldsymbol{x})^T \langle f, \boldsymbol{k}(\cdot)\rangle_{\mathcal{H}_K} \right| \\ &= \left| \langle f, K(\cdot, \boldsymbol{x}) - \ell^T(\boldsymbol{x}) \boldsymbol{k}(\cdot)\rangle_{\mathcal{H}_K} \right| \\ &\leq \|f\|_{\mathcal{H}_K} \left\| K(\cdot, \boldsymbol{x}) - \boldsymbol{k}^T(\boldsymbol{x}) \mathrm{K}^{-1} \boldsymbol{k}(\cdot) \right\|_{\mathcal{H}_K} = \|f\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x}), \end{aligned}$$

with power function

$$P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{K(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1} \boldsymbol{k}(\boldsymbol{x})}.$$

The standard error bound

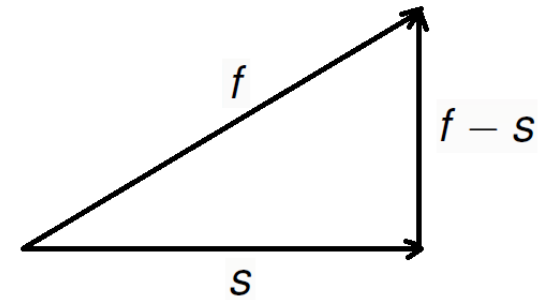
$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \|f\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x})$$

can be improved to the tight bound (see [GW59])

$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \|f - s\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x})$$

since $f - s$ is orthogonal to s in $\mathcal{H}_K$, i.e.,

$$\|f\|^2_{\mathcal{H}_K} = \|f - s\|^2_{\mathcal{H}_K} + \|s\|^2_{\mathcal{H}_K}.$$



The standard error bound

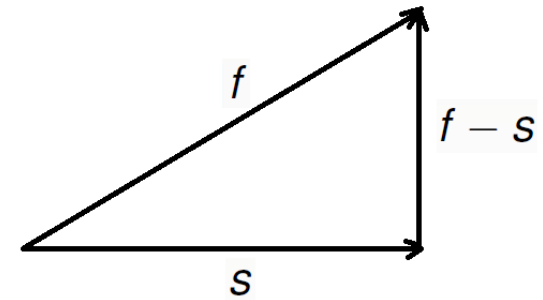
$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \|f\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x})$$

can be improved to the tight bound (see [GW59])

$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \|f - s\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x})$$

since $f - s$ is orthogonal to s in $\mathcal{H}_K$, i.e.,

$$\|f\|^2_{\mathcal{H}_K} = \|f - s\|^2_{\mathcal{H}_K} + \|s\|^2_{\mathcal{H}_K}.$$



Problem: Usually, neither $\|f\|_{\mathcal{H}_K}$ nor $\|f - s\|_{\mathcal{H}_K}$ are computable.

The standard error bound

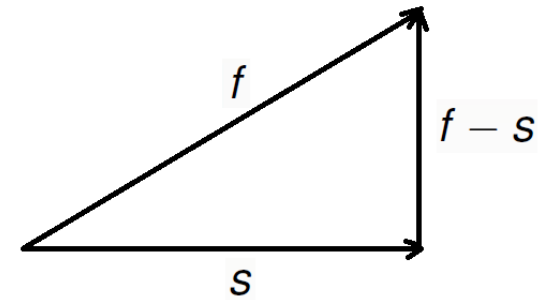
$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \|f\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x})$$

can be improved to the tight bound (see [GW59])

$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \|f - s\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x})$$

since $f - s$ is orthogonal to s in $\mathcal{H}_K$, i.e.,

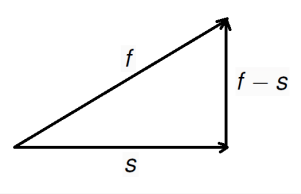
$$\|f\|^2_{\mathcal{H}_K} = \|f - s\|^2_{\mathcal{H}_K} + \|s\|^2_{\mathcal{H}_K}.$$



Problem: Usually, neither $\|f\|_{\mathcal{H}_K}$ nor $\|f - s\|_{\mathcal{H}_K}$ are computable. But $\|s\|_{\mathcal{H}_K}$ is.

If we assume that our approximation s is not too bad, i.e.,

$$\|f - s\|_{\mathcal{H}_K(\Omega)} \leq \delta \|s\|_{\mathcal{H}_K(\Omega)}$$

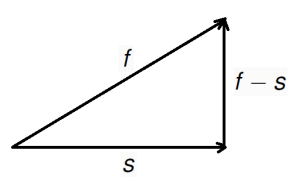


for some not too large constant δ, then the Golomb–Weinberger tight error bound yields a computable error bound

$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \delta \|s\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x}).$$

If we assume that our approximation s is not too bad, i.e.,

$$\|f - s\|_{\mathcal{H}_K(\Omega)} \leq \delta \|s\|_{\mathcal{H}_K(\Omega)}$$



for some not too large constant δ, then the Golomb–Weinberger tight error bound yields a computable error bound

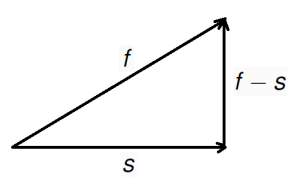
$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \delta \|s\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x}).$$

This is indeed computable since

$$\begin{aligned} \|s\|^2_{\mathcal{H}_K} &= \langle \boldsymbol{y}^T \boldsymbol{\ell}(\cdot), \boldsymbol{\ell}(\cdot)^T \boldsymbol{y} \rangle_{\mathcal{H}_K} = \langle \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{k}(\cdot), \boldsymbol{k}(\cdot)^T \mathrm{K}^{-1} \boldsymbol{y} \rangle_{\mathcal{H}_K} \\ &= \boldsymbol{y}^T \mathrm{K}^{-1} \langle \boldsymbol{k}(\cdot), \boldsymbol{k}(\cdot)^T \rangle_{\mathcal{H}_K} \mathrm{K}^{-1} \boldsymbol{y} = \boldsymbol{y}^T \mathrm{K}^{-1} \mathrm{K} \mathrm{K}^{-1} \boldsymbol{y}. \end{aligned}$$

If we assume that our approximation s is not too bad, i.e.,

$$\|f - s\|_{\mathcal{H}_K(\Omega)} \leq \delta \|s\|_{\mathcal{H}_K(\Omega)}$$



for some not too large constant δ, then the Golomb–Weinberger tight error bound yields a computable error bound

$$|f(\boldsymbol{x}) - s(\boldsymbol{x})| \leq \delta \|s\|_{\mathcal{H}_K} P_{K,\mathcal{X}}(\boldsymbol{x}).$$

This is indeed computable since

$$\begin{aligned}\|s\|^2_{\mathcal{H}_K} &= \langle \boldsymbol{y}^T \boldsymbol{\ell}(\cdot), \boldsymbol{\ell}(\cdot)^T \boldsymbol{y} \rangle_{\mathcal{H}_K} = \langle \boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{k}(\cdot), \boldsymbol{k}(\cdot)^T \mathrm{K}^{-1} \boldsymbol{y} \rangle_{\mathcal{H}_K} \\ &= \boldsymbol{y}^T \mathrm{K}^{-1} \langle \boldsymbol{k}(\cdot), \boldsymbol{k}(\cdot)^T \rangle_{\mathcal{H}_K} \mathrm{K}^{-1} \boldsymbol{y} = \boldsymbol{y}^T \mathrm{K}^{-1} \mathrm{K} \mathrm{K}^{-1} \boldsymbol{y}.\end{aligned}$$

Therefore we have

$$\mathrm{C}_{\mathrm{GW}}(\varepsilon; p) = \sqrt{\boldsymbol{y}^T \mathrm{K}^{-1} \boldsymbol{y}}\, \|P_{K,\mathcal{X}}\|_p,$$

where we compute the p-norm on a discrete evaluation grid.

Kriging Variance

Using the optimal weights $\overset{\star}{\boldsymbol{w}}(\cdot) = \mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x})$, the kriging variance (MSE of the kriging predictor) does depend on the process variance:

$$\begin{aligned}
\mathrm{MSE}(\widehat{Y}_{\boldsymbol{x}}) &= \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \widehat{Y}_{\boldsymbol{x}}\right)^2\right] \\
&= \sigma^2 K(\boldsymbol{x},\boldsymbol{x}) - 2\overset{\star}{\boldsymbol{w}}(\boldsymbol{x})^T(\sigma^2\boldsymbol{k}(\boldsymbol{x})) + \overset{\star}{\boldsymbol{w}}(\boldsymbol{x})^T(\sigma^2\mathrm{K})\overset{\star}{\boldsymbol{w}}(\boldsymbol{x}) \\
&= \sigma^2\left(K(\boldsymbol{x},\boldsymbol{x}) - 2\boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\mathrm{K}\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x})\right) \\
&= \sigma^2\left(K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x})\right).
\end{aligned}$$

Kriging Variance

Using the optimal weights $\overset{\star}{\boldsymbol{w}}(\cdot) = \mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x})$, the kriging variance (MSE of the kriging predictor) does depend on the process variance:

$$\begin{aligned}
\mathrm{MSE}(\widehat{Y}_{\boldsymbol{x}}) &= \mathbb{E}\left[\left(Y_{\boldsymbol{x}} - \widehat{Y}_{\boldsymbol{x}}\right)^2\right] \\
&= \sigma^2 K(\boldsymbol{x},\boldsymbol{x}) - 2\overset{\star}{\boldsymbol{w}}(\boldsymbol{x})^T(\sigma^2\boldsymbol{k}(\boldsymbol{x})) + \overset{\star}{\boldsymbol{w}}(\boldsymbol{x})^T(\sigma^2\mathsf{K})\overset{\star}{\boldsymbol{w}}(\boldsymbol{x}) \\
&= \sigma^2\left(K(\boldsymbol{x},\boldsymbol{x}) - 2\boldsymbol{k}(\boldsymbol{x})^T\mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x})^T\mathsf{K}^{-1}\mathsf{K}\mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x})\right) \\
&= \sigma^2\left(K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathsf{K}^{-1}\boldsymbol{k}(\boldsymbol{x})\right).
\end{aligned}$$

As for the MLE criterion, we can use the optimal profile variance

$$\sigma^2_{\text{opt}} = \frac{1}{N}\boldsymbol{y}^T\mathsf{K}^{-1}\boldsymbol{y}$$

to see that the kriging variance is essentially the same as the Golomb–Weinberger criterion.

Outline

Computing the Cardinal Functions

In the standard basis we find the cardinal basis functions $\ell_j(\boldsymbol{x}_i) = \delta_{ij}$ via

$$\mathsf{K}\boldsymbol{\ell}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x}) \quad \Longleftrightarrow \quad \boldsymbol{\ell}(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T \mathsf{K}^{-1},$$

where $\boldsymbol{\ell}(\cdot) = \begin{pmatrix} \ell_1(\cdot) & \cdots & \ell_N(\cdot) \end{pmatrix}^T$.

Computing the Cardinal Functions

In the standard basis we find the cardinal basis functions $\ell_j(\boldsymbol{x}_i) = \delta_{ij}$ via

$$\mathsf{K}\boldsymbol{\ell}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x}) \quad \Longleftrightarrow \quad \boldsymbol{\ell}(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T\mathsf{K}^{-1},$$

where $\boldsymbol{\ell}(\cdot) = \begin{pmatrix}\ell_1(\cdot) & \cdots & \ell_N(\cdot)\end{pmatrix}^T$.

Moreover, in any alternate basis defined by (see [PS11])

$$\boldsymbol{v}(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T\mathsf{T}, \qquad \mathsf{K} = \mathsf{V}\mathsf{T}^{-1},$$

we have

$$\begin{aligned}\boldsymbol{\ell}(\boldsymbol{x})^T &= \boldsymbol{k}(\boldsymbol{x})^T\mathsf{K}^{-1} \\ &= \boldsymbol{v}(\boldsymbol{x})^T\mathsf{T}^{-1}\mathsf{T}\mathsf{V}^{-1} \\ &= \boldsymbol{v}(\boldsymbol{x})^T\mathsf{V}^{-1}.\end{aligned}$$

Examples of Alternate Bases

Newton basis [MS09, PS11]: via Cholesky factorization $\mathsf{K} = \mathsf{N}\mathsf{N}^T$

$$\boldsymbol{n}(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T \mathsf{N}^{-T}, \quad \mathsf{V} = \mathsf{N}, \quad \mathsf{T} = \mathsf{N}^{-T}$$

SVD basis [PS11, DMS13]: via SVD $\mathsf{K} = \mathsf{Q}\Sigma^2\mathsf{Q}^T$

$$\boldsymbol{v}(\boldsymbol{x})^T = \boldsymbol{k}(\boldsymbol{x})^T \mathsf{Q}\Sigma^{-1}, \quad \mathsf{V} = \mathsf{Q}\Sigma, \quad \mathsf{T} = \mathsf{Q}\Sigma^{-1}$$

HS-SVD basis [FM12]: via HS-SVD $\Psi\Lambda_1\Phi_1^T = \mathsf{K}$

$$\psi(\boldsymbol{x})^T = \phi(\boldsymbol{x})^T \begin{pmatrix} \mathsf{I}_N \\ \Lambda_2\Phi_2^T\Phi_1^{-T}\Lambda_1^{-1} \end{pmatrix}, \quad \mathsf{V} = \Psi, \quad \mathsf{T} = \Phi_1^{-T}\Lambda_1^{-1}$$

Remark

Can also use low rank approximate bases (see, e.g., [PS11, FM12]).

Computing log det K

We compute determinants using logarithms to prevent underflow errors that may/will arise for small enough values of the shape parameter.

Computing log det K

We compute determinants using logarithms to prevent underflow errors that may/will arise for small enough values of the shape parameter.

Mathematically, computing log det K is straightforward and stable using the Hilbert–Schmidt SVD $K = \Psi \Lambda_1 \Phi_1^T$, i.e.,

$$\log\det K = \log\det \Psi + \log\det \Lambda_1 + \log\det \Phi_1^T.$$

Computing log det K

We compute determinants using logarithms to prevent underflow errors that may/will arise for small enough values of the shape parameter.

Mathematically, computing log det K is straightforward and stable using the Hilbert–Schmidt SVD $K = \Psi \Lambda_1 \Phi_1^T$, i.e.,

$$\log\det K = \log\det \Psi + \log\det \Lambda_1 + \log\det \Phi_1^T.$$

Computationally,

- the very small eigenvalues can be handled safely by taking their logarithms (since Λ_1 is diagonal),
- Φ_1^T gets inverted while forming the stable basis, and
- Ψ gets inverted while computing an interpolant, so the cost of computing log(det(K)) is negligible.

Computing the Power Function

In addition to the ill-conditioning which may be present in the matrix K (and so K^{-1}), evaluation of the power function is susceptible to numerical cancelation.

Computing the Power Function

In addition to the ill-conditioning which may be present in the matrix K (and so K^{-1}), evaluation of the power function is susceptible to numerical cancelation.

Consider

$$\widetilde{\mathrm{K}} = \begin{pmatrix} K(\boldsymbol{x},\boldsymbol{x}) & \boldsymbol{k}(\boldsymbol{x})^T \\ \boldsymbol{k}(\boldsymbol{x}) & \mathrm{K} \end{pmatrix},$$

so that

$$\det(\widetilde{\mathrm{K}}) = \det\left(\begin{pmatrix} 1 & \boldsymbol{k}(\boldsymbol{x})^T \\ \mathbf{0}_N & \mathrm{K} \end{pmatrix}\begin{pmatrix} K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) & \mathbf{0}_N^T \\ \mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) & \mathrm{I}_N \end{pmatrix}\right)$$

Computing the Power Function

In addition to the ill-conditioning which may be present in the matrix K (and so K^{-1}), evaluation of the power function is susceptible to numerical cancelation.

Consider

$$\widetilde{\mathrm{K}} = \begin{pmatrix} K(\boldsymbol{x},\boldsymbol{x}) & \boldsymbol{k}(\boldsymbol{x})^T \\ \boldsymbol{k}(\boldsymbol{x}) & \mathrm{K} \end{pmatrix},$$

so that

$$\begin{aligned} \det(\widetilde{\mathrm{K}}) &= \det\left(\begin{pmatrix} 1 & \boldsymbol{k}(\boldsymbol{x})^T \\ \boldsymbol{0}_N & \mathrm{K} \end{pmatrix} \begin{pmatrix} K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) & \boldsymbol{0}_N^T \\ \mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) & \mathrm{I}_N \end{pmatrix} \right) \\ &= \det(\mathrm{K})(K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x})). \end{aligned}$$

Computing the Power Function

In addition to the ill-conditioning which may be present in the matrix K (and so K^{-1}), evaluation of the power function is susceptible to numerical cancelation.

Consider

$$\widetilde{\mathrm{K}} = \begin{pmatrix} K(\boldsymbol{x},\boldsymbol{x}) & \boldsymbol{k}(\boldsymbol{x})^T \\ \boldsymbol{k}(\boldsymbol{x}) & \mathrm{K} \end{pmatrix},$$

so that

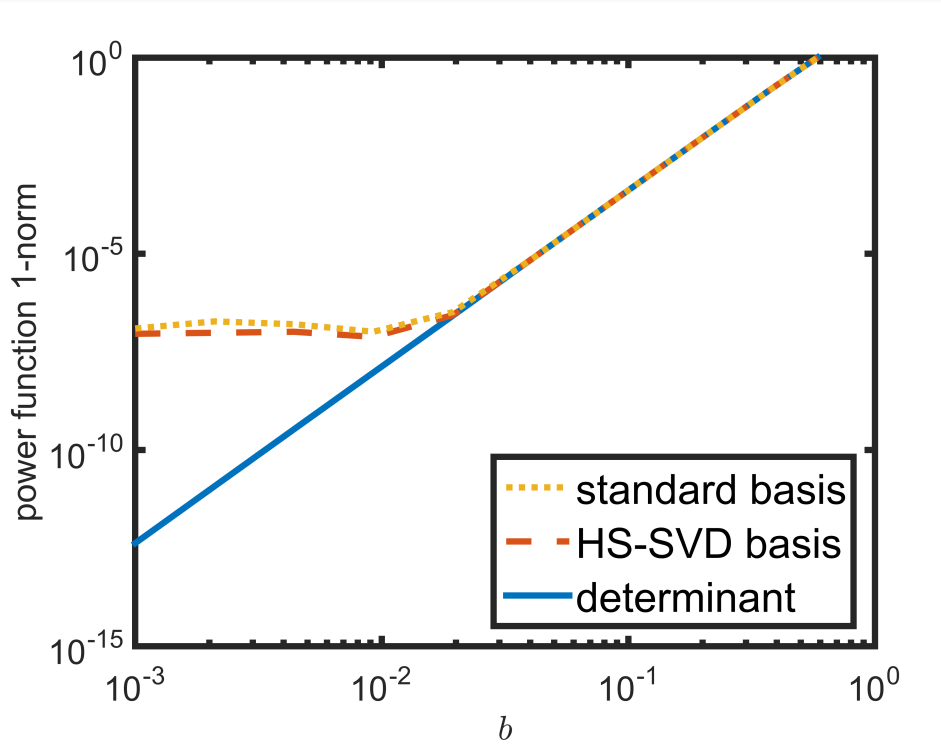
$$\begin{aligned} \det(\widetilde{\mathrm{K}}) &= \det\left(\begin{pmatrix} 1 & \boldsymbol{k}(\boldsymbol{x})^T \\ \boldsymbol{0}_N & \mathrm{K} \end{pmatrix} \begin{pmatrix} K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) & \boldsymbol{0}_N^T \\ \mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) & \mathrm{I}_N \end{pmatrix} \right) \\ &= \det(\mathrm{K})(K(\boldsymbol{x},\boldsymbol{x}) - \boldsymbol{k}(\boldsymbol{x})^T\mathrm{K}^{-1}\boldsymbol{k}(\boldsymbol{x})). \end{aligned}$$

Therefore, the power function can be computed via [FWL04, Sch05]

$$P_{K,\mathcal{X}}(\boldsymbol{x}) = \sqrt{\frac{\det(\widetilde{\mathrm{K}})}{\det(\mathrm{K})}}.$$

Outline

Computing the Power Function — Example

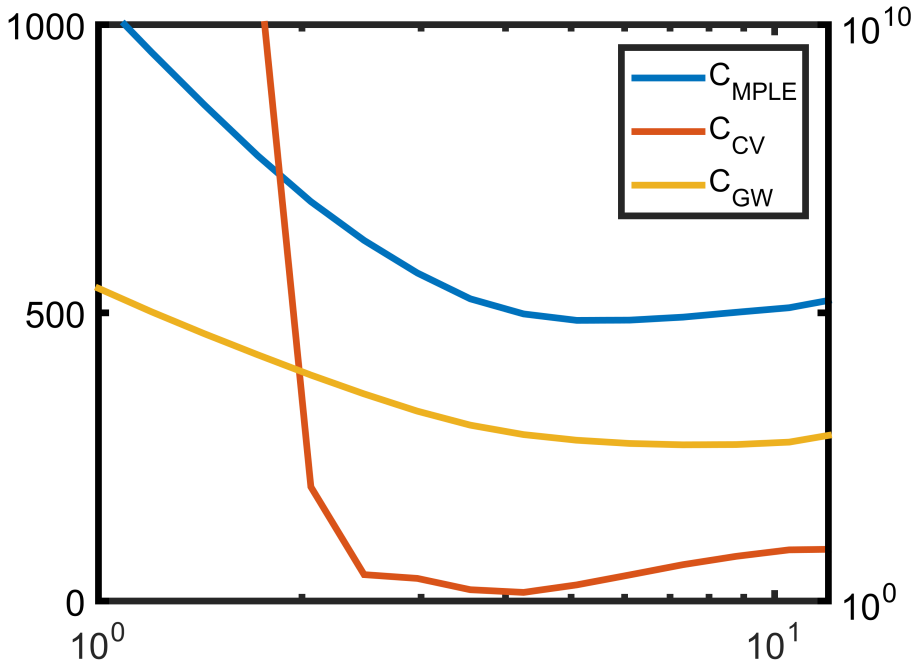


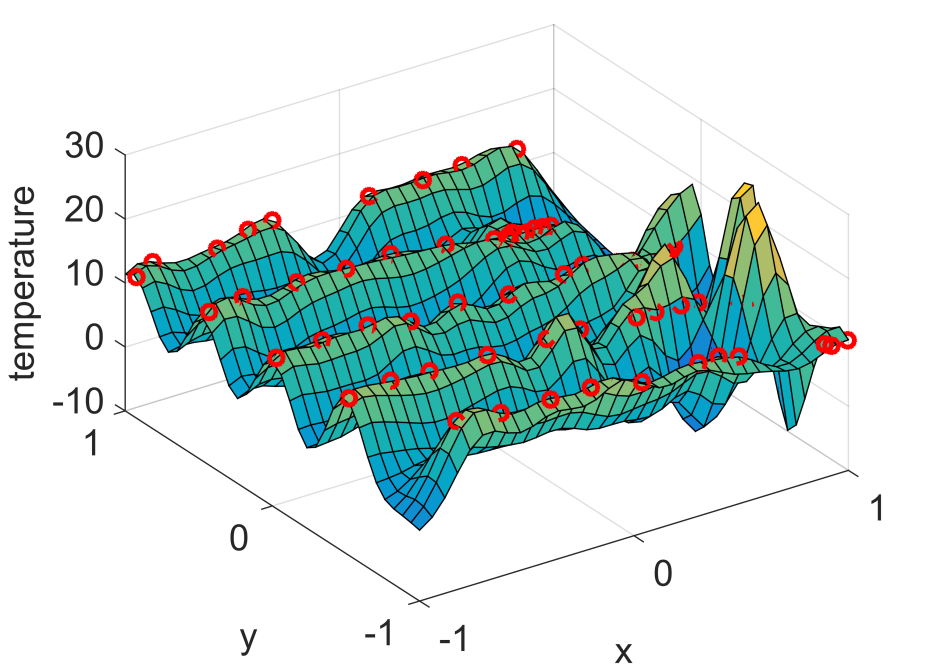
Analytic Chebyshev kernel $K(x,z) = \sum_{n=0}^{\infty} \lambda_n \varphi_n(x)\varphi_n(z)$ on 11 Chebyshev points in $[-1,1]$

$$\lambda_0 = \frac{1}{2}, \quad \lambda_n = \frac{(1-b)b^n}{2b}, \quad \varphi_n(x) = \sqrt{2-\delta_{n0}}\,T_n(x),$$

$$K(x,z) = \frac{1}{2} + (1-b)\frac{b(1-b^2) - 2b(x^2+z^2) + (1+3b^2)xz}{(1-b^2)^2 + 4b\left(b(x^2+z^2) - (1+b^2)xz\right)}$$

Using various criteria and isotropic kernels to fit track data [Dav14]



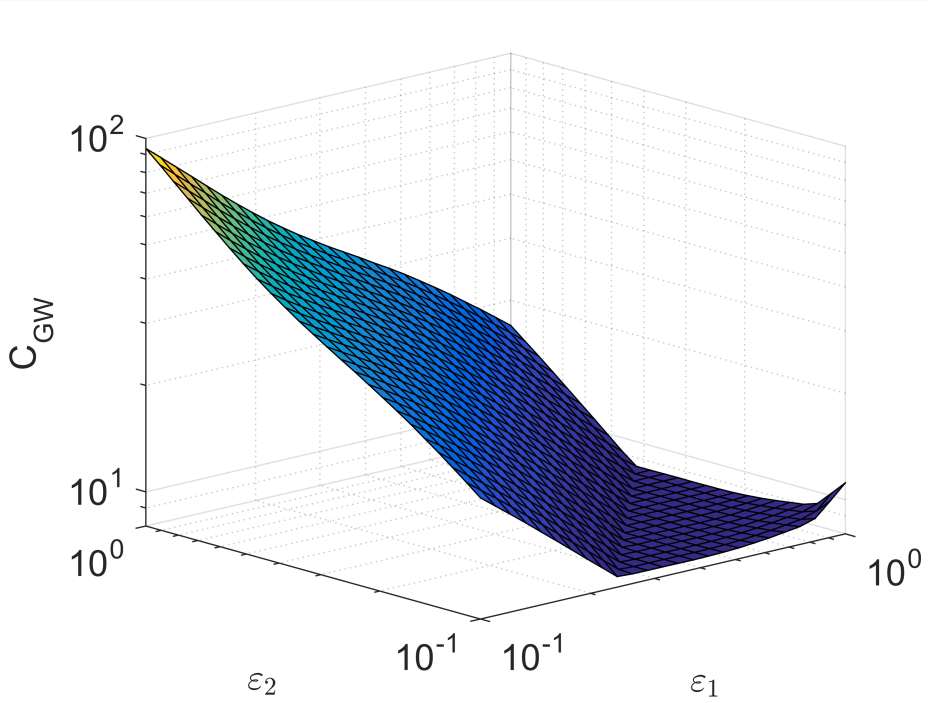


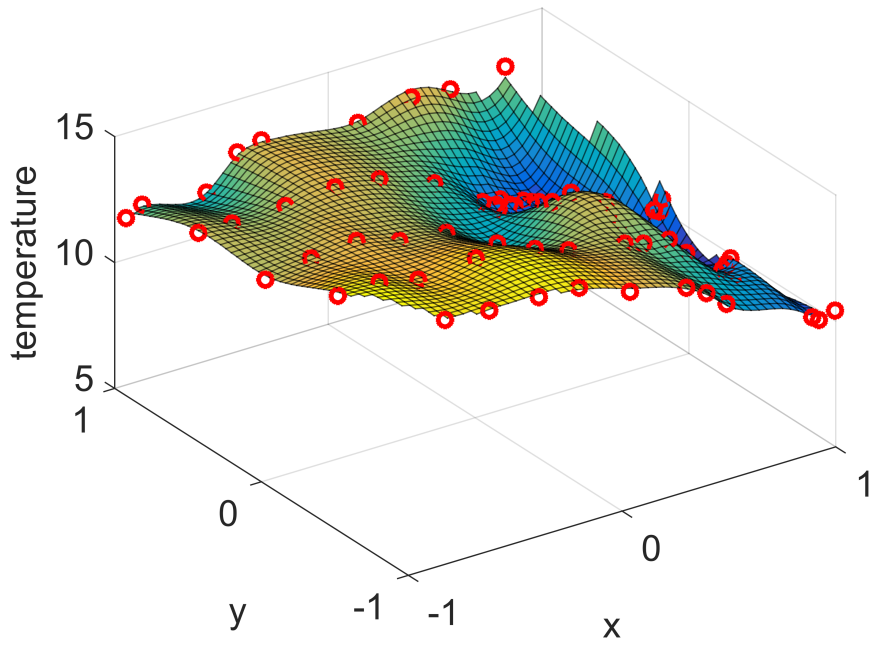
Interpolation with isotropic Gaussian kernel, $\varepsilon = 6$:

$$\varepsilon_{\text{GWopt}} = 7.3162, \quad \varepsilon_{\text{CVopt}} = 4.2518, \quad \varepsilon_{\text{MPLEopt}} = 5.0950$$

Using C_{GW} and anisotropic kernels to fit track data

[Dav14]



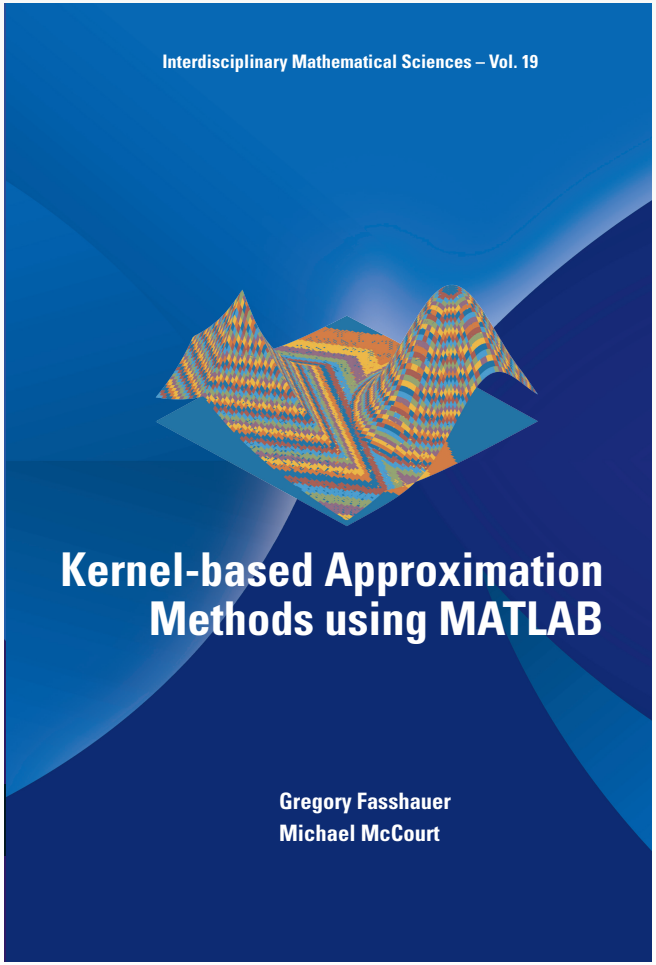


Interpolation with anisotropic tensor product C^2 Matérn kernels:

$$\varepsilon_{1\text{opt}} = 2.37\varepsilon_{2\text{opt}}, \quad \varepsilon_{1\text{opt}} = 0.6803, \varepsilon_{2\text{opt}} = 0.2875$$

Summary

- Explained various criteria for choosing "optimal" kernel parameters (including C_{GW}, based on error bound)
- Reliable evaluation of these criteria requires
  - alternate (stable) bases
  - kriging variance with process variance
  - determinant formula for power function
- Vast applications
  - function interpolation/approximation
  - numerical solution of PDEs (collocation, MFS, MPS, RBF-FD)
  - machine learning (RBF network regression, low-rank approximation for SVM classification)
  - ...



MATLAB code available at

`http://math.iit.edu/~mccomic/gaussqr`

One more inspiring perspective

Calanque

References I

[Dav14] Oleg Davydov, *Tsfit: A software package for two-stage scattered data fitting*, https://www.staff.uni-giessen.de/odavydov/tsfit/, February 2014.

[DMS13] Stefano De Marchi and Gabriele Santin, *A new stable basis for radial basis function interpolation*, J. Comput. Appl. Math. **253** (2013), 1–13.

[FM12] G. E. Fasshauer and M. J. McCourt, *Stable evaluation of Gaussian radial basis function interpolants*, SIAM J. Sci. Comput. **34** (2012), no. 2, A737–A762.

[FWL04] B. Fornberg, G. Wright, and E. Larsson, *Some observations regarding interpolants in the limit of flat radial basis functions*, Comput. Math. Appl. **47** (2004), 37–55.

[GW59] M. Golomb and H. F. Weinberger, *Optimal approximation and error bounds*, On Numerical Approximation (R. E. Langer, ed.), University of Wisconsin Press, 1959, pp. 117–190.

[MS09] Stefan Müller and Robert Schaback, *A Newton basis for kernel spaces*, J. Approx. Theory **161** (2009), no. 2, 645–655.

References II

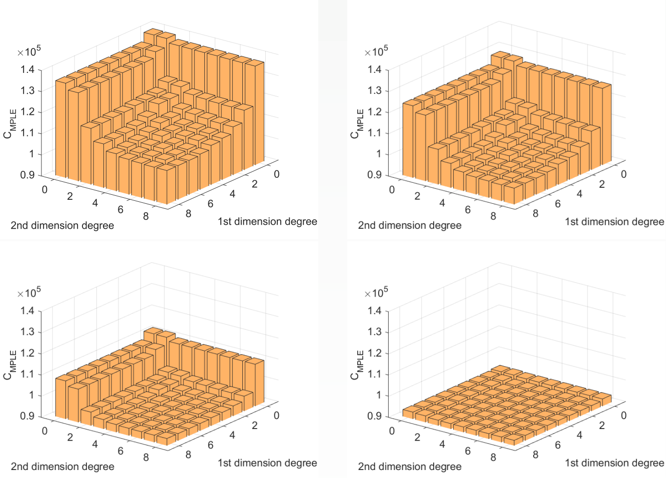
[PS11] M. Pazouki and R. Schaback, *Bases for kernel-based spaces*, J. Comput. Appl. Math. **236** (2011), no. 4, 575–588.

[Sch05] Robert Schaback, *Multivariate interpolation by polynomials and radial basis functions*, Constr. Approx. **21** (2005), 293–317.

[Sch11] _____, *The missing Wendland functions*, Adv. Comput. Math. **34** (2011), no. 1, 67–81.

Using C_{MPLE} with universal kriging to fit glacier data

[Dav14]



Interpolation with anisotropic Wendland kernels and polynomials:

$$\varepsilon = (20\ 21), \quad \varepsilon = (16\ 14), \quad \varepsilon = (8\ 11), \quad \varepsilon = (5\ 4)$$

Using CMPLE with universal kriging to fit glacier data

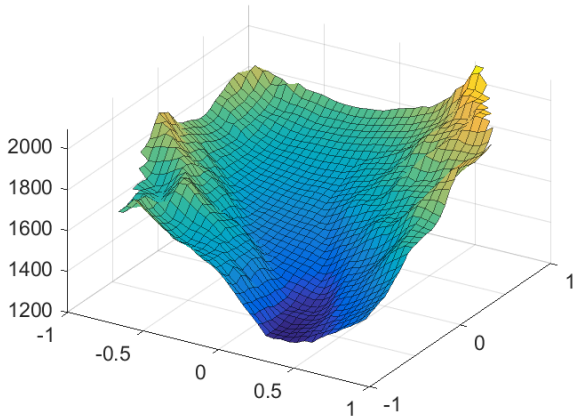
[Dav14]

Interpolation with anisotropic "missing" Wendland kernels [Sch11]

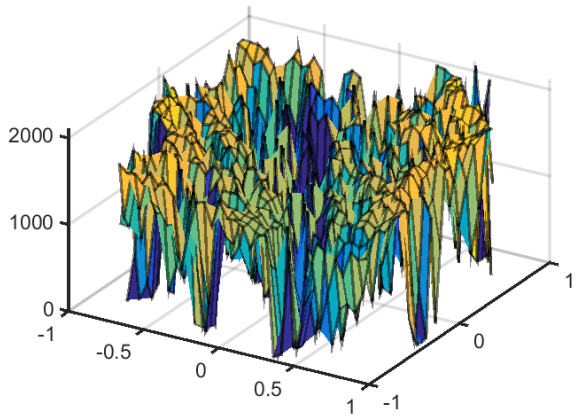
$$\kappa_{3,\frac{3}{2}}(r) \doteq \left(1 - 7r^2 - \frac{81}{4}r^4\right)\sqrt{1-r^2} - \frac{15}{4}r^4(6+r^2)\log\left(\frac{r}{1+\sqrt{1-r^2}}\right)$$

and polynomial trend

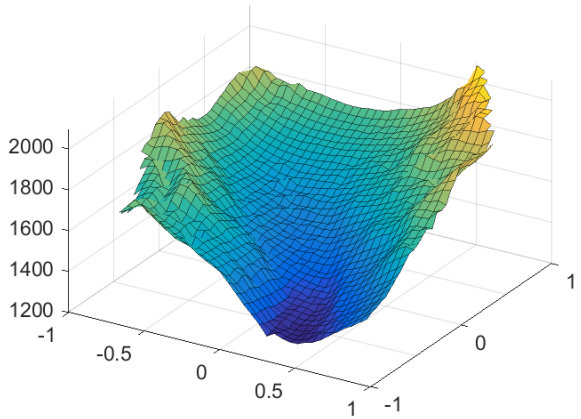
$$s(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^T \mathrm{K}^{-1}(\boldsymbol{y} - \mathrm{P}\beta) + \boldsymbol{p}(\boldsymbol{x})^T\beta, \quad \beta = (\mathrm{P}^T\mathrm{K}^{-1}\mathrm{P})^{-1}\mathrm{P}^T\mathrm{K}^{-1}\boldsymbol{y}.$$



$\varepsilon = (5\ 4), \deg = (0\ 0)$



$\varepsilon = (20\ 21)$



$\varepsilon = (20\ 21), \deg = (8\ 8)$

COLORADO SCHOOL OF MINES

Improved efficiency with hybrid/multiscale methods

ε	degree	density	times (s) $K = LL^T$	$C_{MPLE}(\varepsilon, \boldsymbol{p})$	eval
(20 21)	0	0.25%	0.59	0.03	0.32
(20 21)	8		0.59	0.21	0.33
(5 4)	0	4.17%	4.04	0.35	0.63
(5 4)	8		4.04	1.45	0.64

Handle

- large-scale trends with polynomials (or global kernels)
- small-scale features with compactly supported kernels.