Kriging and Alternatives in Computer Experiments

C. F. Jeff Wu
ISyE, Georgia Institute of Technology

- Use kriging to build meta-models in computer experiments, a brief review
- Numerical problems with kriging
- Alternatives to kriging:
  - Regularized kriging, Hybrid kriging
  - Overcomplete basis surrogate model (OBSM)

Georgia Institute of Technology

ISyE

Why computer experiments?

- ✔ No need for expensive lab equipments and materials, less costly than physical experiments.

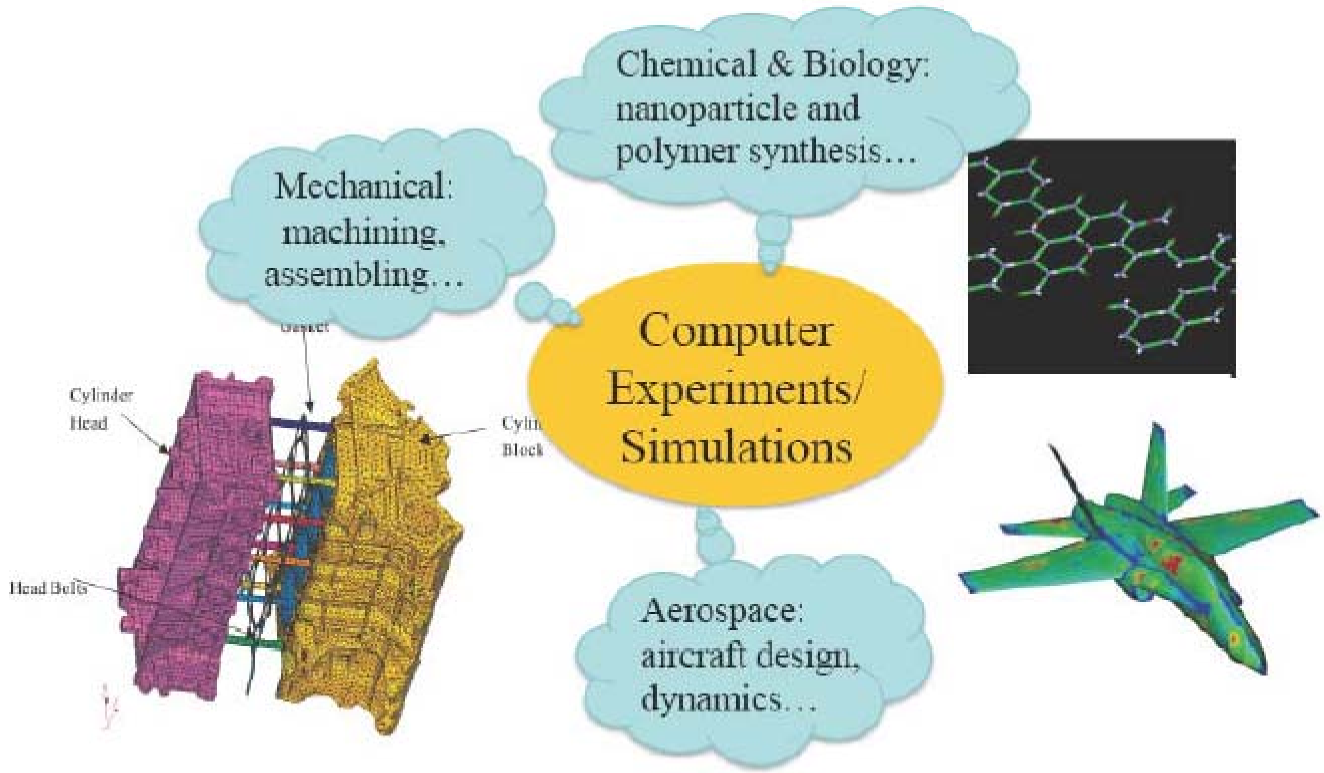
- ✔ Not affected by human and environmental factors.

- ✔ Study dangerous or infeasible physical experiments, such as ammunition detonation.

Georgia Institute of Technology

ISyE

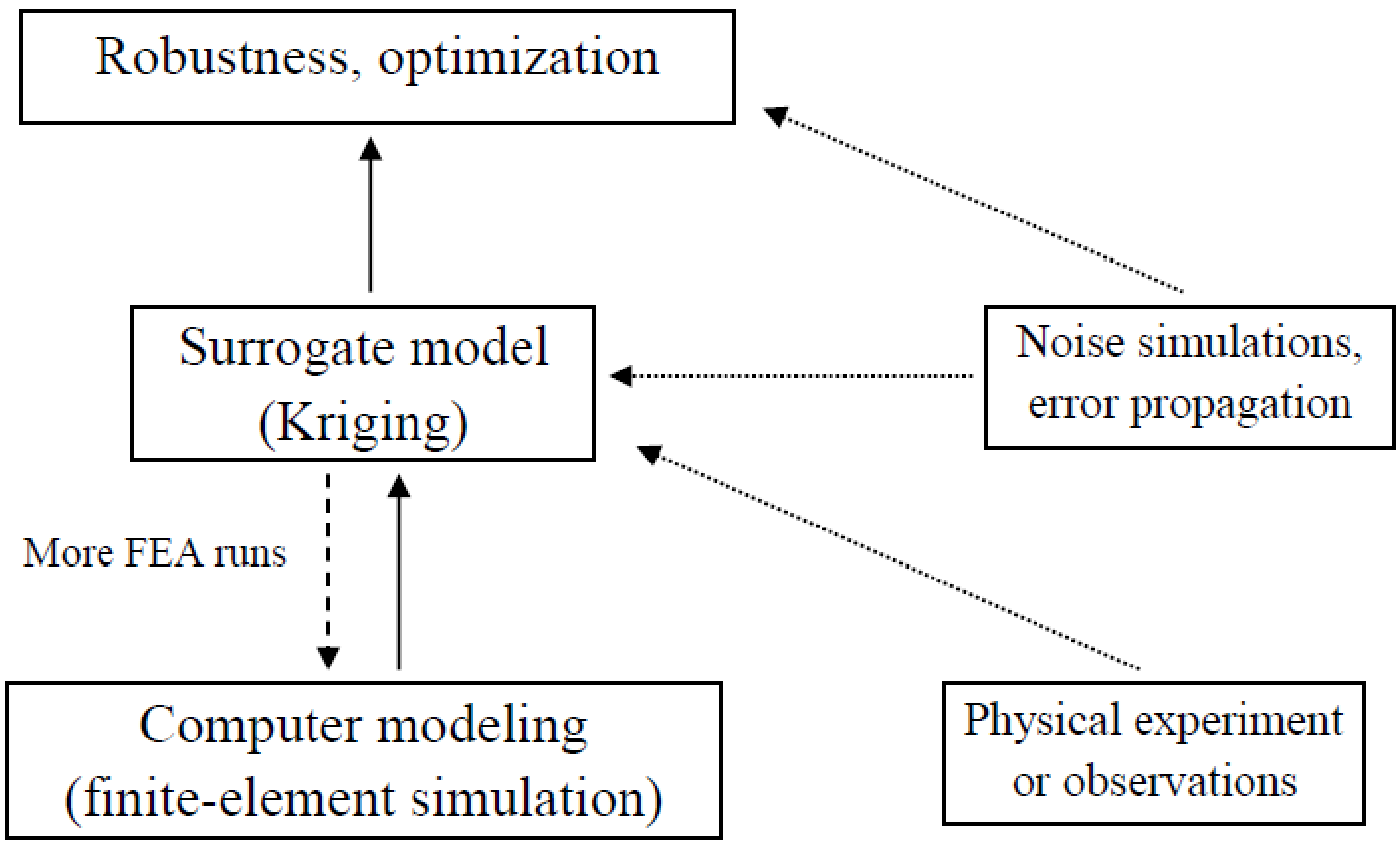
Some examples



Georgia Institute of Technology

ISyE

Statistical Meta-Modeling of Computer Experiments



Georgia Institute of Technology

ISyE

Kriging Models

- Ordinary Kriging

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x})$$

$$Z(\mathbf{x}) \sim N_n\left(\mathbf{0},\ \sigma^2\varphi(\mathbf{h})\right) \equiv GP\left(\mathbf{0},\ \sigma^2\varphi(\mathbf{h})\right)$$

- Correlation function
  - $\varphi(\mathbf{0}) = 1$
  - $\varphi(\mathbf{h}) = \varphi(-\mathbf{h})$, (symmetric function)
  - φ is a positive semi-definite function

Georgia Institute of Technology

ISyE

Correlation function

- Matérn

$$\varphi(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(2\sqrt{\nu}\theta|h|\right)^{\nu} K_{\nu}\left(2\sqrt{\nu}\theta|h|\right)$$

where K_ν is the modified Bessel function of order ν

$$\nu \to \infty, \quad \varphi(h) \to \exp(-\theta h^2)$$

- ***Power exponential correlation***

$$\varphi(h) = \exp\left(-\theta|h|^{q}\right), \quad 0 < q \le 2, \quad 0 < \theta$$

  - q =2 ***Gaussian correlation function*** **(infinitely differentiate)**
  - q =1 Ornstein-Uhlenbeck process (ν = 1 in Matern)

- Linear, Cubic correlation

Kriging predictor

- Best Linear Unbiased Predictor (BLUP)

$$\hat{y}(\boldsymbol{x}) = \hat{\mu} + r(\boldsymbol{x})'\boldsymbol{R}^{-1}(\boldsymbol{y} - \hat{\mu}\mathbf{1}),$$

$$r(\boldsymbol{x})' = (\varphi(\boldsymbol{x} - \boldsymbol{x}_1), \ldots, \varphi(\boldsymbol{x} - \boldsymbol{x}_n)),$$

$$\hat{\mu} = \mathbf{1}'\boldsymbol{R}^{-1}\boldsymbol{y}/\mathbf{1}'\boldsymbol{R}^{-1}\mathbf{1},$$

$\widehat{y}(\boldsymbol{x}_i) = y_i$ *an* interpolating property*

*required for deterministic simulations

Georgia Institute of Technology ISyE

Recent work in kriging

- Calibration of computer model, Kriging with *calibration* parameters (Kennedy-O'Hagan, 2001), with *tuning* parameters (Santner et al., 2009)
- Computer simulations with different levels of accuracy (Kennedy-O'Hagan, 2000; Qian et al., 2006; Qian-Wu, 2008) ➡

  construction of nested space-filling (e.g., Latin hypercube) designs (Qian-Ai-Wu, 2009, various papers by Qian and others, 2009-date)

Georgia Institute of Technology

ISyE

Recent work in kriging (cont.)

- Kriging for multiple outputs and functional response (Conti et al., 2009; Conti and O'Hagan, 2010)
- Treed Gaussian Process model (Gramacy and Lee, 2008).
- Kriging (i.e., GP model) with quantitative and qualitative factors (Qian-Wu-Wu, 2008, Han et al., 2009) ⇒

  construction of sliced space-filling (e.g., Latin hypercube) designs (Qian-Wu, 2009, Qian, 2010)

Georgia Institute of Technology ISyE

Maximum Likelihood Estimation

- Profile log-likelihood approach

$$Q(\boldsymbol{\theta}) = n\log(\sigma^2(\boldsymbol{\theta})) + \log|\boldsymbol{R}(\boldsymbol{\theta})|$$

$$\text{where } \sigma^2(\boldsymbol{\theta}) = \{\boldsymbol{y} - \hat{\mu}(\boldsymbol{\theta})\mathbf{1}\}'\boldsymbol{R}^{-1}(\boldsymbol{\theta})\{\boldsymbol{y} - \hat{\mu}(\boldsymbol{\theta})\mathbf{1}\}/n$$

$$\hat{\mu}(\boldsymbol{\theta}) = \mathbf{1}'\boldsymbol{R}^{-1}(\boldsymbol{\theta})\boldsymbol{y}/\mathbf{1}'\boldsymbol{R}^{-1}(\boldsymbol{\theta})\mathbf{1}$$

Numerical Instability in $R^{-1}(\theta)$

- $R(\theta)$ is an *n*x*n* matrix, *n*=sample size
- Its condition number (max e.v./min e.v.) ↑ as

I. Sample size *n* ↑

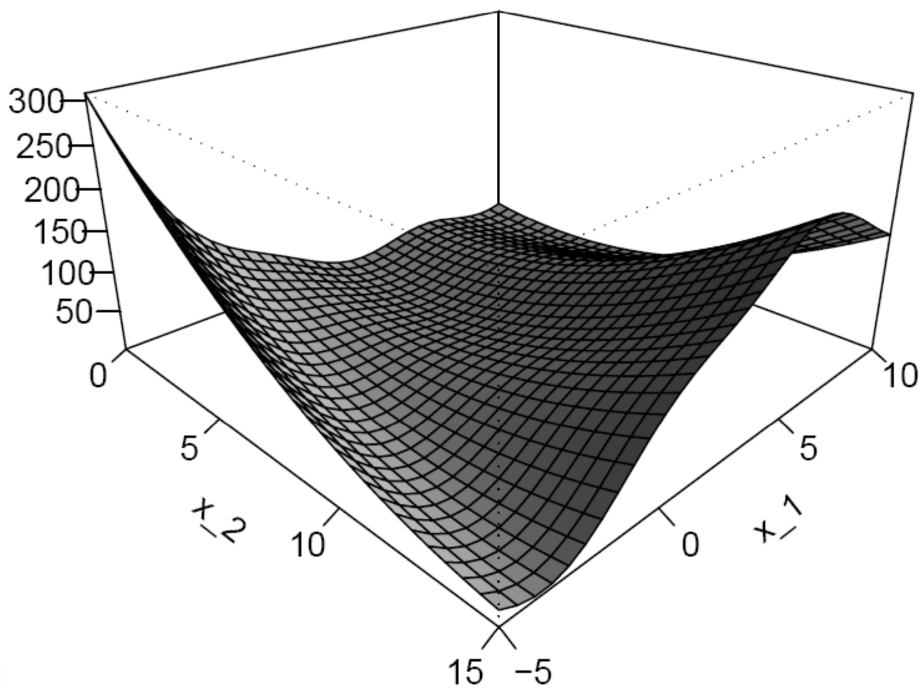
II. Dimension of input vectors ↑

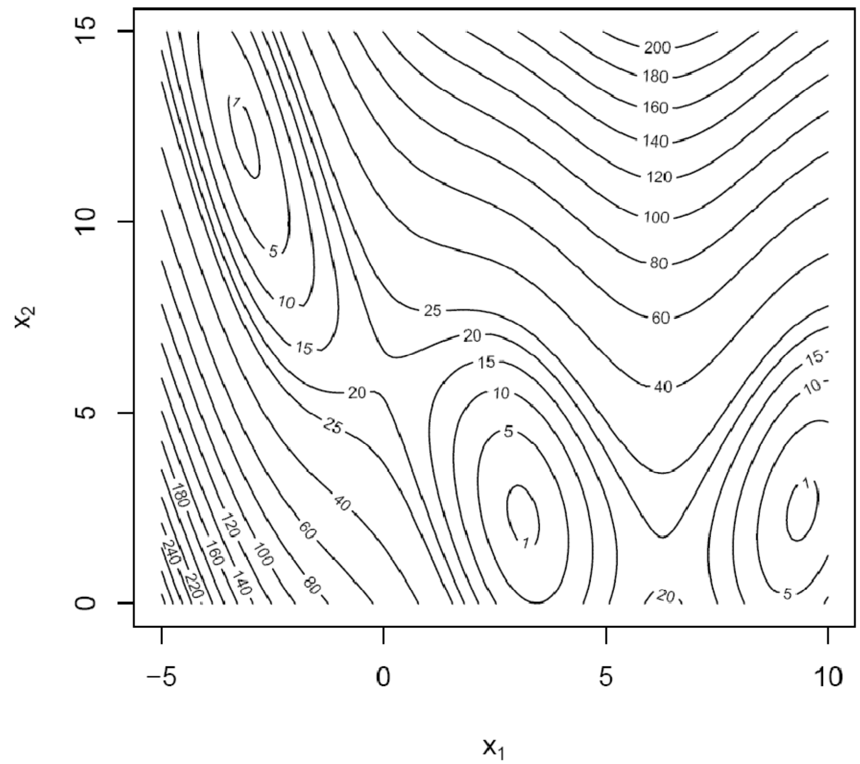
(Peng-Wu, 2010)

Branin function

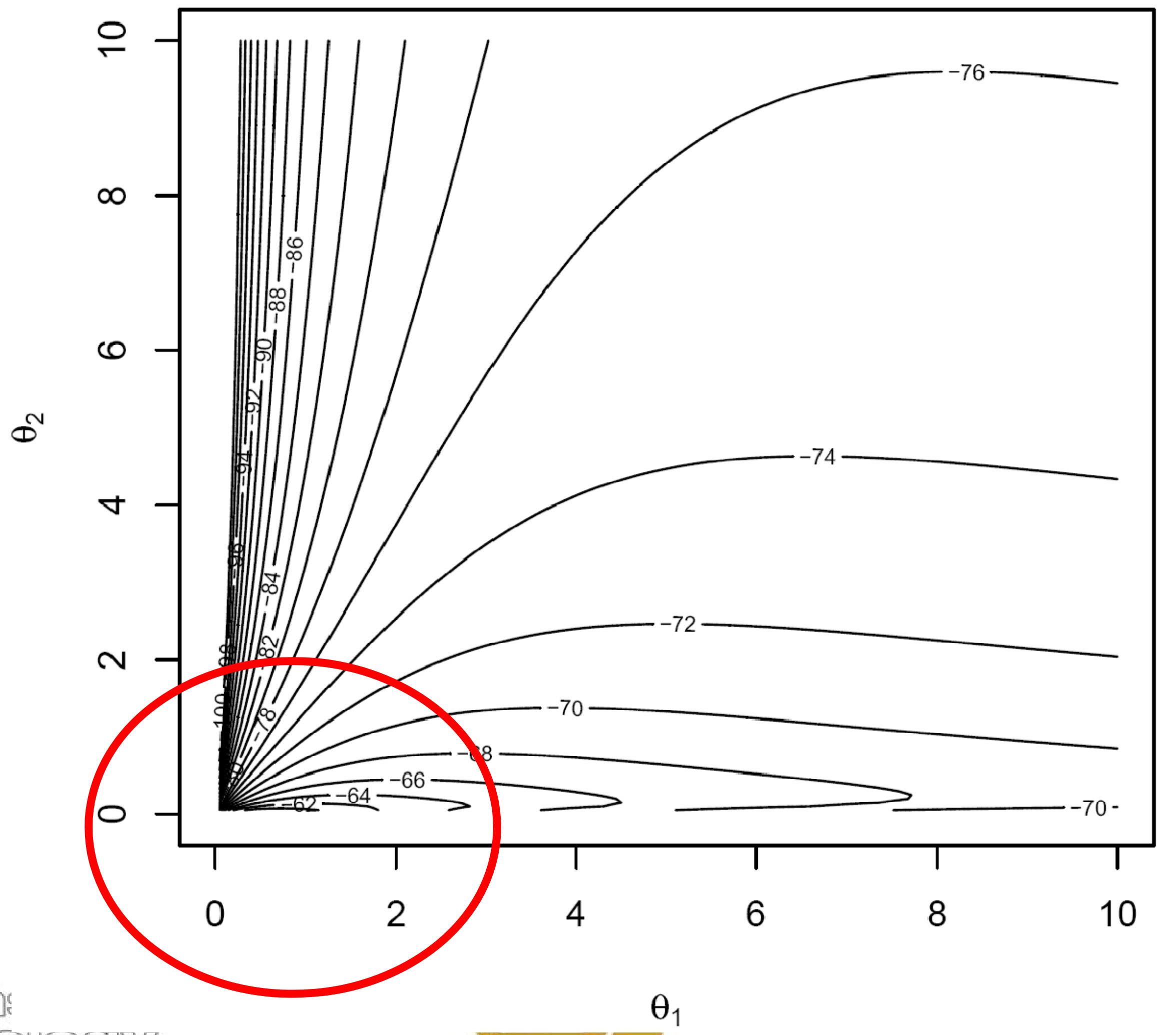
(Andre, Siarry and Dognon, 2001)

$$f(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10$$





Log-likelihood function
(Regular grid: $n = 7^2$)



Regularized Kriging

- Introducing a regularizing constant λ into the predictor

$$\hat{y}_\lambda(\boldsymbol{x}) = \hat{\mu}_\lambda + r(\boldsymbol{x})'(\boldsymbol{R} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{y} - \hat{\mu}_\lambda \mathbf{1})$$

$$\text{where } \hat{\mu}_\lambda = \mathbf{1}'(\boldsymbol{R} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} / \mathbf{1}'(\boldsymbol{R} + \lambda \boldsymbol{I})^{-1}\mathbf{1}$$

  Peng and Wu (2010, submitted)

- Similar modification in estimation: maximizing a **regularized likelihood**

Kriging with nugget effects

- Model from spatial statistics

$$Y(\boldsymbol{x}) = \mu + Z(\boldsymbol{x}) + \delta\boldsymbol{\epsilon}$$

- BLUP

$$\hat{y}_\delta(\boldsymbol{x}) = \hat{\mu}_\delta + r(\boldsymbol{x})'\left(\boldsymbol{R} + \frac{\delta^2}{\sigma^2}\boldsymbol{I}\right)^{-1}(\boldsymbol{y} - \hat{\mu}_\delta\mathbf{1})$$

$$\hat{\mu}_\delta = \frac{\mathbf{1}'\left(\boldsymbol{R} + \frac{\delta^2}{\sigma^2}\boldsymbol{I}\right)^{-1}\boldsymbol{y}}{\mathbf{1}'\left(\boldsymbol{R} + \frac{\delta^2}{\sigma^2}\boldsymbol{I}\right)^{-1}\mathbf{1}}$$

Algorithm (Ridge Trace)

- Root Mean Squared Prediction Error (RMSPE)

$$\text{RMSPE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y(\mathbf{x}_i) - \hat{Y}(\mathbf{x}_i))^2}$$

(1) Set λ^* as the lower bound and choose a grid point set for λ, say, $(\lambda_1, \ldots, \lambda_k)$, and let $i = 1$.

(2) Use λ_i in regularized kriging to estimate $\boldsymbol{\theta}_\lambda$.

(3) Compute the RMSPE. Let $i = i + 1$.

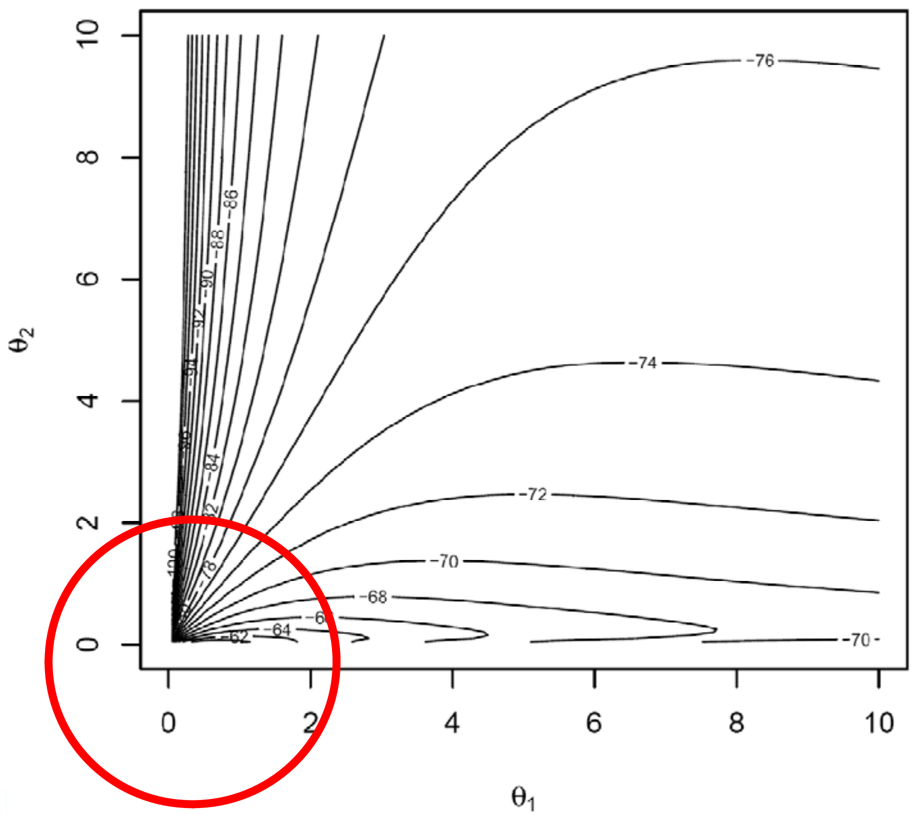
(4) Repeat steps 2 and 3 until all k grid points are exhausted.

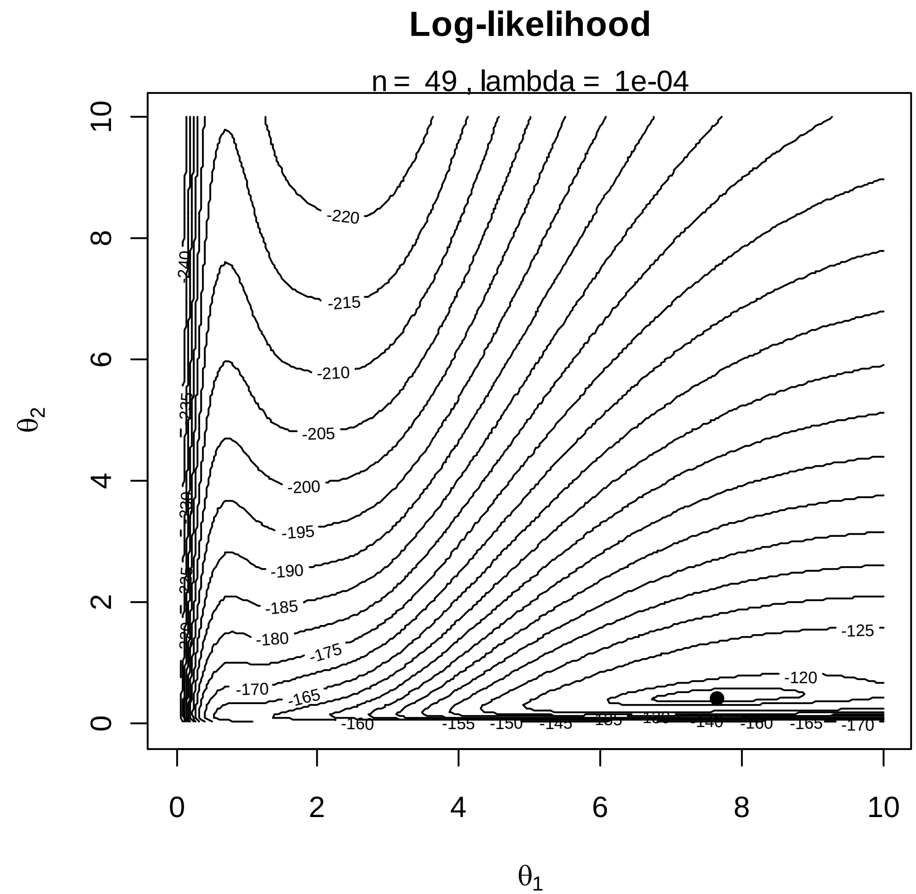
(5) The final estimator $\hat{\boldsymbol{\theta}}_{\hat{\lambda}}$ is the one with the lowest RMSPE with $\hat{\lambda}$.

Georgia Institute of Technology

ISyE

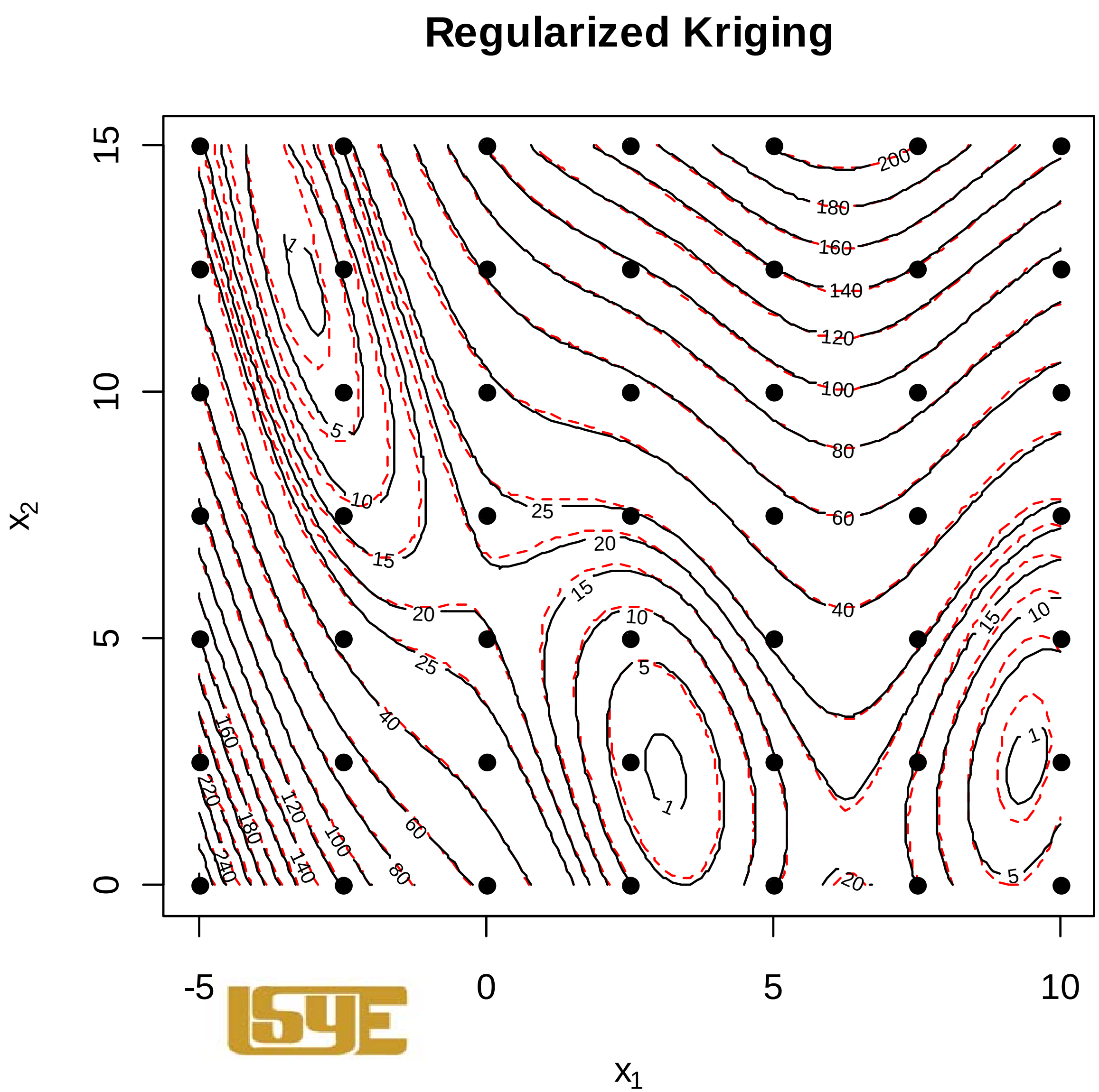
Log-likelihood function, Branin function, (Regular grid: $n = 7^2$)





$$\lambda^* = \frac{n^{1/2}}{\Delta^{1/2} - 1} = 3.3 \times 10^{-7}$$

λ	RMSPE
10^{-1}	6.1842
10^{-2}	3.2085
10^{-3}	1.5605
10^{-4}	1.0706
10^{-5}	1.7069
10^{-6}	2.5104
10^{-7}	3.2348



Georgia Institute of Technology

Overcomplete Basis Surrogate Model

- Use an overcomplete dictionary of basis functions
- Use linear combinations of basis functions to approximate unknown functions
- Use Matching Pursuit for fast (i.e. greedy) computations
- Choice of basis functions to “mimic” the shape of the surface

Chen, Wang, and Wu (2010, *IIE Tran*. Q&R)

Surrogate Representation

- Surrogate model: use a linear combination of pre-specified basis functions, i.e.,

$$f\,(\boldsymbol{x}) = \sum_j c_j\ \phi_j(\boldsymbol{x})\ ,\ \boldsymbol{x} \in \chi$$

  – unitary norm $\|\phi_j\| = 1$
  – basis dictionary, $\{\phi_j, j = 1, \ldots, M\}$
  – no unknown parameter in ϕ_j , only unknown are the *linear* c_j
- Overcomplete : M much larger than data size

Georgia Institute of Technology

ISYE

Surrogate Model (continued)

- Explored point set: $P_{exp} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p\}$.
- Current responses:

$$V_{P\exp} = \left(f(x_1), \ldots, f(x_p)\right)^{\mathrm{T}}$$

- Use $\sum_j c_j \tilde{\phi}_j$ to approximate $V_{P\exp}$,

$$\tilde{\phi}_j = \left(\phi_j(x_1), \ldots, \phi_j(x_p)\right)^{\mathrm{T}}$$

- Two interesting questions:
  - *Choice of* the basis functions?
  - Estimation of the linear coefficients c_i?

Coefficient Inference

- Matching Pursuit Algorithm (Mallat and Zhang, 1993):
  - Infer coefficients by minimizing $\| V_{P\exp} - \sum_j c_j \tilde{\phi}_j \|$
  - A greedy algorithm: at the *j*th iteration,

    Let $R^{(j-1)}$ be the current residual vector.

    Selected a basis by $\tilde{\phi}_{(j)} = \arg\max_i \left\langle R^{(j-1)}, \tilde{\phi}_i \right\rangle$:

$$c_{(j)} = c_{(j)} + \left\langle R^{(j-1)}, \tilde{\phi}_{(j)} \right\rangle,$$

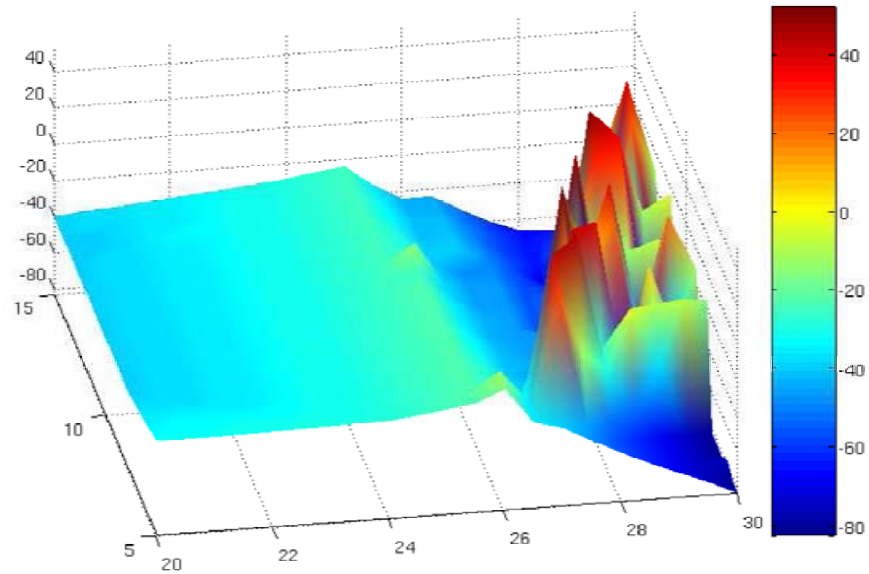
$$R^{(j)} = R^{(j-1)} - \left\langle R^{(j-1)}, \tilde{\phi}_{(j)} \right\rangle \tilde{\phi}_{(j)}.$$

Georgia Institute of Technology

ISyE

Response Surface for Bistable Laser Diodes

- The true surface over a pre-specified grid:



- Search all positive Lyapunov exponents (PLE) (red area)
- PLE corresponds to chaotic light output.

Gabor Functions

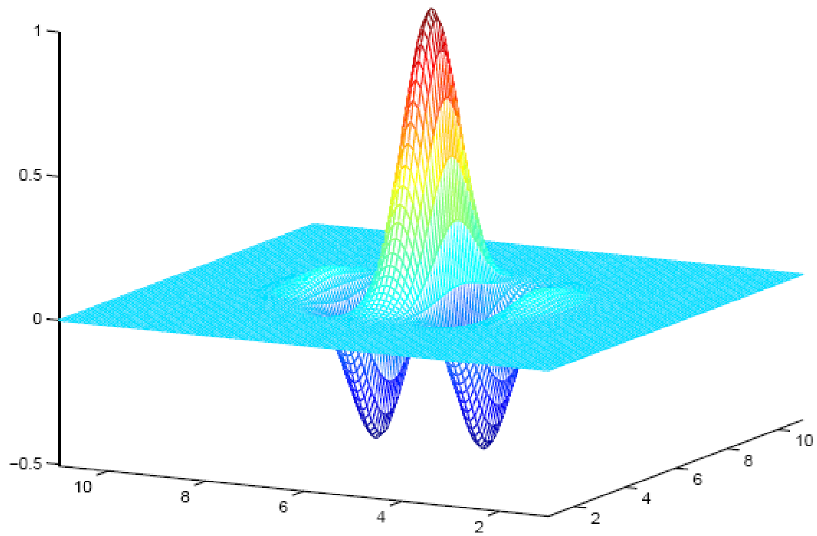
- Basis functions:
  - *n*-dimensional Gabor function

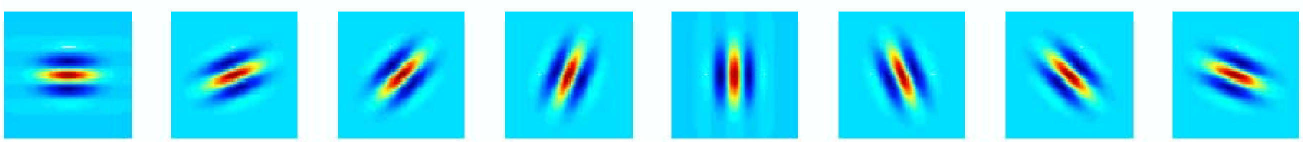
$$g(\boldsymbol{x}) = \exp(-\frac{\boldsymbol{x}^{\mathsf{T}} M \boldsymbol{x}}{2}) \exp(2\pi i A \boldsymbol{x}), \; \boldsymbol{x} = (x_1, \ldots, x_n)^{\mathsf{T}}$$

  - Two-dimensional Gabor function, i.e. *n* = 2

$$\begin{aligned} g(u, v) &= \frac{1}{Z} \exp\left[-\frac{1}{2}(\sigma_u u^2 + \sigma_v v^2)\right] \cos\left[\frac{2\pi u}{\lambda} + \varphi\right], \\ u &= u_0 + x_1 \cos\theta - x_2 \sin\theta \\ v &= v_0 + x_1 \sin\theta - x_2 \cos\theta \end{aligned}$$

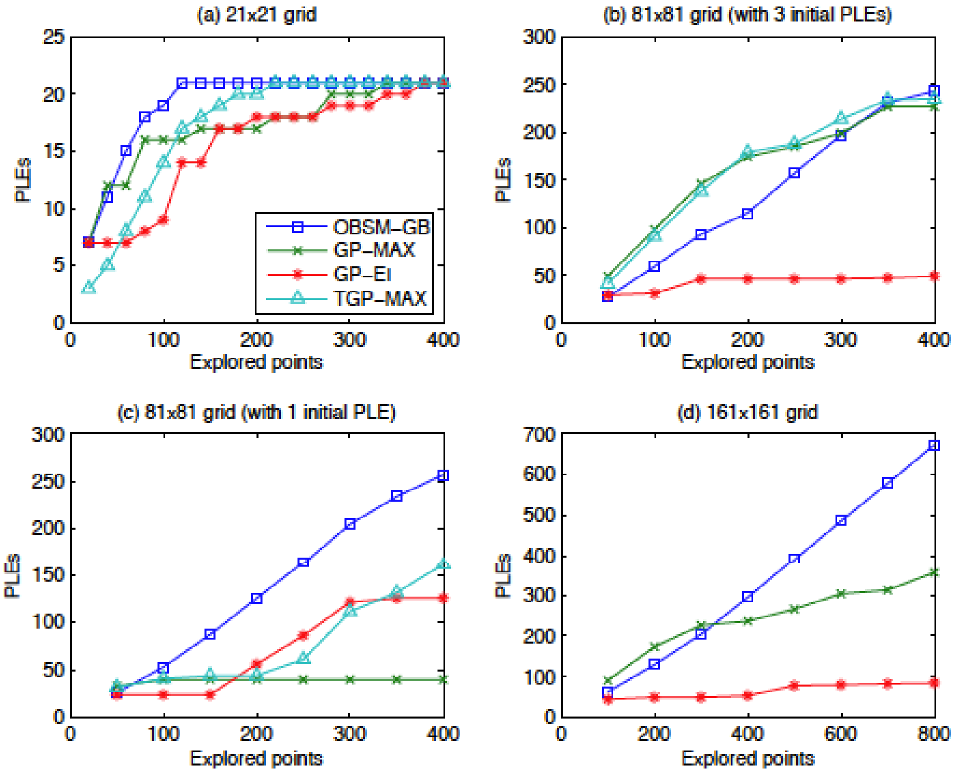
Plots of 2-D Gabor Function





Overall Comparisons

Figure 6: Cumulative numbers of PLEs found by using different explored points in 21×21, 81×81, and 161×161 grids.



Summary

- Computer experiments/simulations have become popular in engineering and science
- Kriging is the most common method for statistical meta-model building but is more limited for large or complex problems
- Alternatives to kriging are being sought:
  - Tweaking of kriging to achieve stability (regularized, hybrid, tapering, reduced rank)
  - Approximations with fast computations (OBSM, RIDW): but *lacking inferential capability*

Covariance Matrix Tapering

- Covariance tapering (Kaufman et al., 2008)

  ✓ Covariance matrix is "tapered" or multiplied element wise by a sparse matrix, to approximate the likelihood.

- Advantages:

  ✓ Significant computational gains/stability.

  ✓ Retain interpolating property.

  ✓ Asymptotic convergence of the tamper estimator.

- But:

  ✓The tapering function is isotropic: OK for spatial statistic problems, but not applicable to engineering problems.

  ✓ The tapering radius needs to be determined.

Rank Reduction

- Fixed rank kriging (Cressie-Johannesson, 2008)
  - ✓ A flexible family of non-stationary covariance function is defined by using a set of basis functions that are fixed in number (smaller than the data size *n*).
- Advantage:
  - ✓ Reduce the computational cost of kriging to $O(n)$.
- But:
  - ✓How to choose the appropriate basis functions.
  - ✓Not an interpolator.

Upper bound

- Upper bound

$$\kappa_2^p(\boldsymbol{R}(\boldsymbol{\theta}) + \lambda \boldsymbol{I}_n; \boldsymbol{X}) \leq \prod_{k=1}^{p} \kappa_2^1(\boldsymbol{R}(0) + \lambda \boldsymbol{I}_{n_k}; D_k)$$

$$\exp(-\theta_k) = 1 \Leftrightarrow \theta_k = 0$$

- The worst case of a correlation matrix

Georgia Institute of Technology

ISyE

Inverse Distance Weighting (IDW)

- Inverse Distance Weighting (Shepard, 1968):

$$\hat{y}(\boldsymbol{x}) = \frac{\sum_{k=1}^{n} w_k(\boldsymbol{x}) y_k}{\sum_{i=1}^{n} w_i(\boldsymbol{x})}.$$

  - $w_i(\boldsymbol{x}) = 1/d(\boldsymbol{x}, \boldsymbol{x}_i)^2.$
  - $d(\boldsymbol{x}, \boldsymbol{x}_i) = \left\{\sum_{j=1}^{p} (x_j - x_{i,j})^2\right\}^{1/2}.$
  - Simple computation but poor prediction.

Regression-Based Inverse Distance Weighting (RIDW)

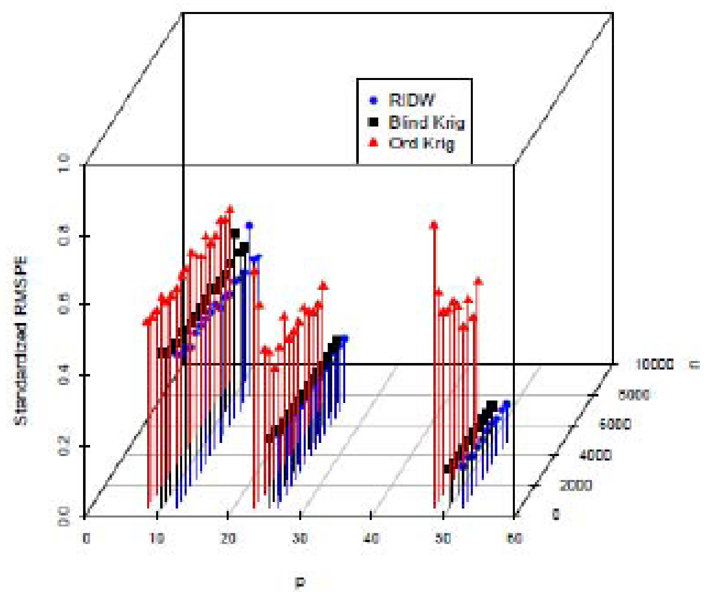
- Add regression part to IDW (Joseph and Kang, 2009):

$$\hat{y}(\boldsymbol{x}) = \mu(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\sum_{k=1}^{n} w_k(\boldsymbol{x}) e_k}{\sum_{i=1}^{n} w_i(\boldsymbol{x})}$$

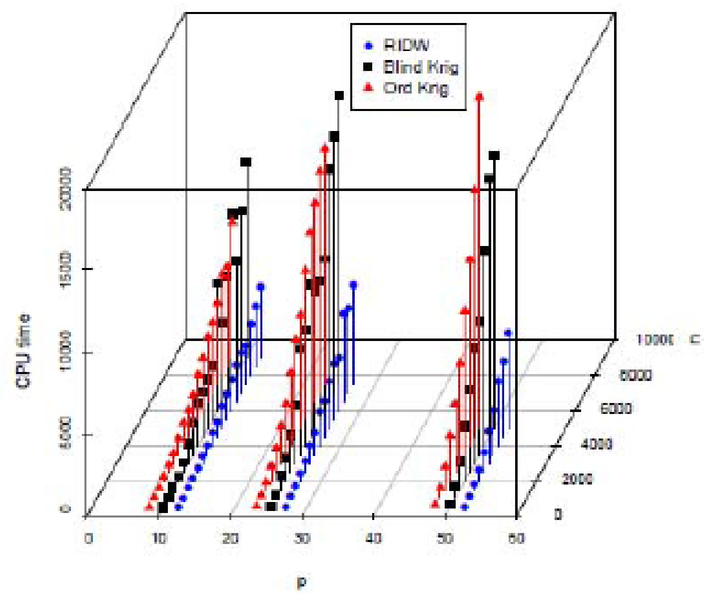
  - $\mu(\boldsymbol{x}_k;\boldsymbol{\beta})$ = mean part; can be linear, nonlinear, nonparametric.
  - $e_k = y_k - \mu(\boldsymbol{x}_k;\boldsymbol{\beta}) = y_k - \mu_k.$
  - $w_i(\boldsymbol{x}) = \frac{\exp\{-d^2(\boldsymbol{x},\boldsymbol{x}_i)\}}{d^2(\boldsymbol{x},\boldsymbol{x}_i)}.$ (faster convergence than IDW)
  - $d(\boldsymbol{x},\boldsymbol{x}_i) = \sqrt{\sum_{j=1}^{p} \theta_j (x_j - x_{i,j})^2}.$

Comparisons Between RIDW and Kriging

Standardized RMSPE



CPU time in simulation



Lower bound on λ

- Lower bound

$$\lambda^* = \inf \left\{ \lambda \,\middle|\, \prod_{k=1}^{p} (1 + n_k/\lambda) < \Delta \right\}$$

where

$$\epsilon = 2^{-52} \approx 2.22 \times 10^{-16}$$

$$\Delta = 1/(10\epsilon) = 4.5 \times 10^{14}$$

Machine accuracy

(or unit round-off)

Georgia Institute of Technology

ISyE