McCormick
Northwestern Engineering
Department of Industrial Engineering and Management Sciences

Better Simulation Metamodeling: The Why, What and How of Stochastic Kriging

Jeremy Staum

Collaborators:
Bruce Ankenman, Barry Nelson
Evren Baysal, Ming Liu, Wei Xie

supported by the NSF under Grant No. DMI-0900354

Winter Simulation WSC'09 Conference

Outline

- overview of metamodeling
- metamodeling approaches
  - regression, kriging, stochastic kriging
- Why kriging? Why stochastic kriging?
- stochastic kriging: what and how
- practical advice for stochastic kriging

Metamodeling: Why?

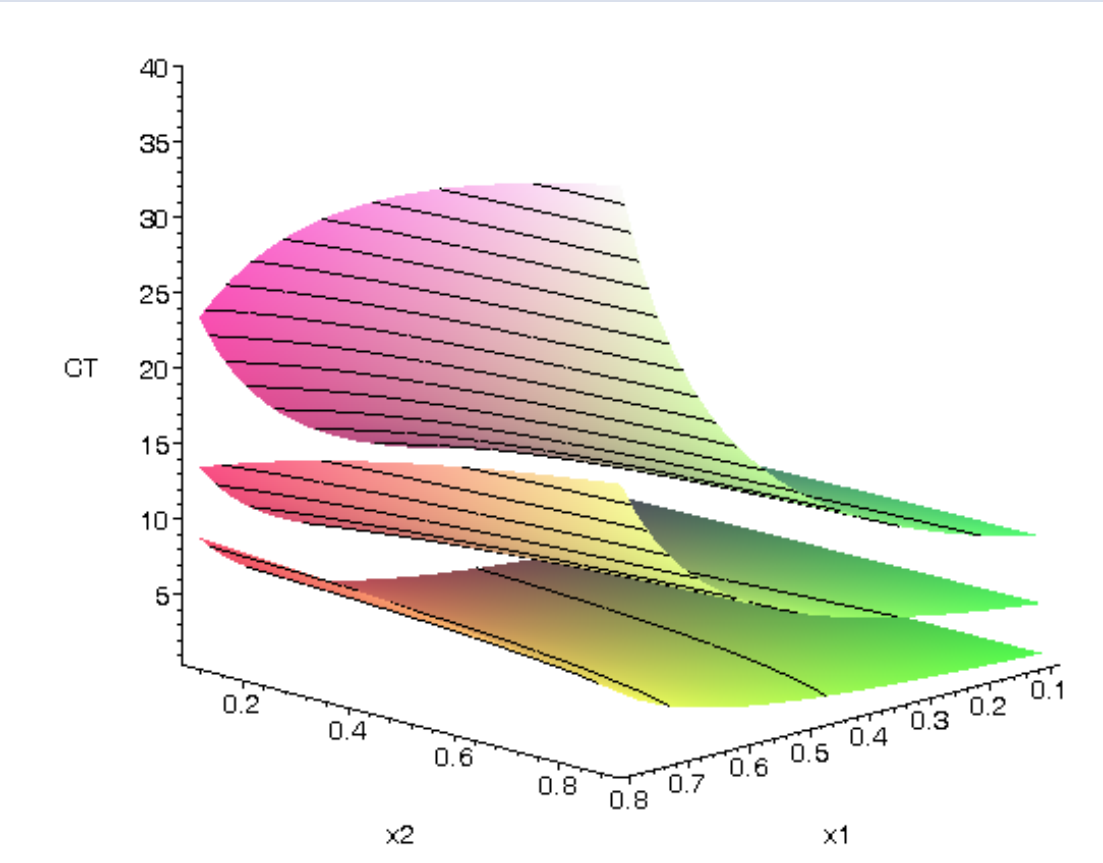
- simulation model
  - input x, response surface y(x)
  - simulation effort n
  - simulation output Y(x;n) estimates y(x)
  - Y(x;n) converges to y(x) as $n \to \infty$
- simulation metamodeling
  - fast estimate of y(x) for any x
  - "What would the simulation say if I ran it?"

Uses of Metamodeling

- **trend modeling** (global)
  - Does $y(x) = y(x_1,x_2)$ increase with x_1?
  - Is $y(x) = y(x_1,x_2)$ more sensitive to x_1 or x_2?
  - $y(x)$ is similar to $\beta_0 + x\beta$ globally
- **optimization** (local)
  - Which way to move and how far?
  - quadratic: first and second derivatives,
    $y(x) \approx \beta_0 + x\beta + x^TQx$ locally

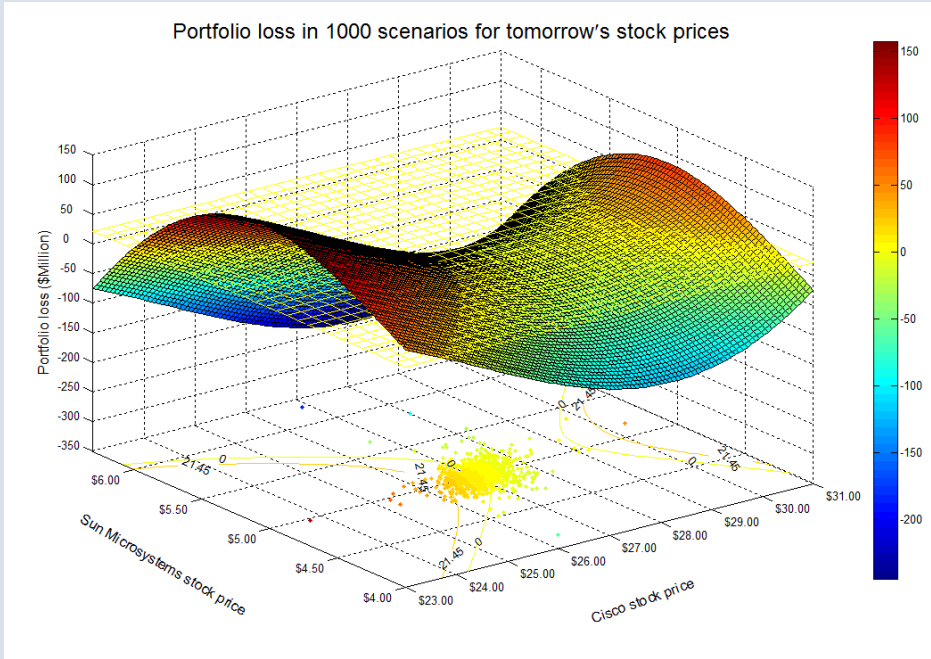
Uses of Metamodeling

- **exploration**
- (global)
- What if?
  scenario = x
- multi-objective tradeoffs
  - throughput (x_1)
  - cycle time (y)



Uses of Metamodeling

- **scenario analysis**
  - (global)
  - can't control scenario
  - financial scenario
  - military scenario
  - simulation inputs
- **probabilistic analysis**
  - distribution on scenarios



Needs of Metamodeling

- **trend modeling:** rough global trend
- **optimization:** rough local trend
- **exploration / scenario analysis:**
  - globally accurate prediction: $\hat{y}(x) \approx y(x)$
  - $\hat{y}(x)$ is almost as good an estimator of $y(x)$ as the simulation output $Y(x;n)$
  - but metamodel is much faster than model
  - “simulation on demand”

Overview of Approaches

- No free lunch!
- Inference about y(x) without seeing Y(x;n) requires assumptions:
  - about spatial variability in y(·)
  - about noise ε(x;n) = Y(x;n) – y(x)
- Is y(·) a simple trend y(x)=b(x)β, or must we model deviation from trend?
- Should we try to filter out noise?

Regression

- Assume $y(x) = b(x)\beta$.
  - The truth y can't deviate from the trend.
  - We aim only to estimate β.
- Assume $\varepsilon(x) = Y(x) - y(x)$ is white noise.
  - Aim to filter it out!
  - Var[$\varepsilon(x)$] doesn't depend on x. (remedies)
- Global metamodeling: can't find an adequate trend model (b).

Interpolation

- Assume $y(\cdot)$ has some smoothness.
  - model mean β_0 and deviation $y(\cdot) - \beta_0$ or
  - trend $b(\cdot)\beta$ and deviation $y(\cdot) - b(\cdot)\beta$
- Assume $\varepsilon(x) = Y(x) - y(x) = 0$.
  - No filtering: $\hat{y}(x) = Y(x)$ if x is simulated.
- Stochastic simulation: need big simulation effort n so $Y(x;n) - y(x) \approx 0$.
  - Interpolated $\hat{y}(\cdot)$ will look bumpy.

Smoothing

- Assume $y(\cdot)$ has some smoothness.
  - just like interpolation
- Assume $\varepsilon(x) = Y(x) - y(x)$ is white noise.
  - Aim to filter it out!
- Can handle ordinal & categorical data.
  - regression, interpolation, or smoothing

Example: M/M/1 Queue

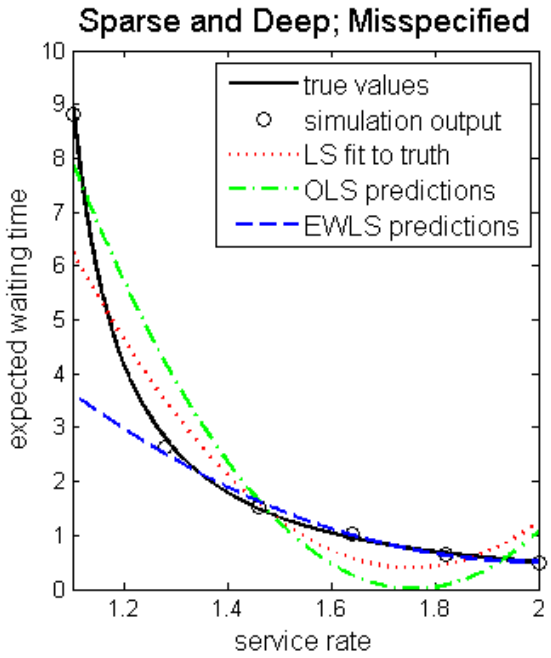
- arrival rate 1
- service rate x
- steady-state mean wait y(x)=1/(x(x-1))
- initialize in steady-state (no bias) and simulate a fixed # of customers
- variance also explodes as x↓1

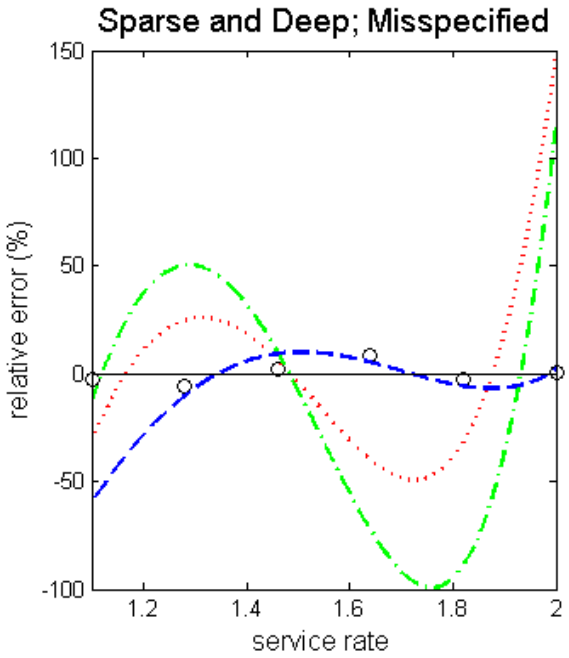
Regression Remedies

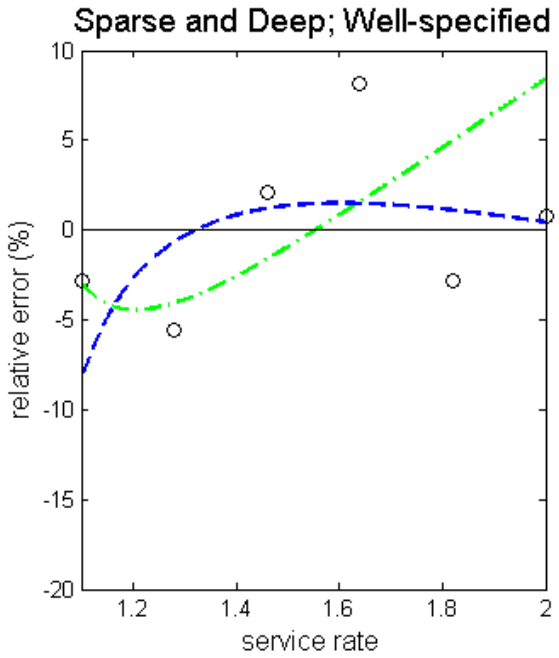
- weighted least squares (WLS)
  - weight on Y(x;n) is n/Var[Y(x)]
  - empirical WLS: estimate Var[Y(x)]
- generalized linear models:
  - $Y(x) = y(x) + \varepsilon(x)$, $y(x) = LINK(b(x)\beta)$
  - perfect for M/M/1: $y(x) = \exp(\beta_0 + \beta_1 \ln(x) + \beta_2 \ln(x-1)) = 1/(x(x-1))$

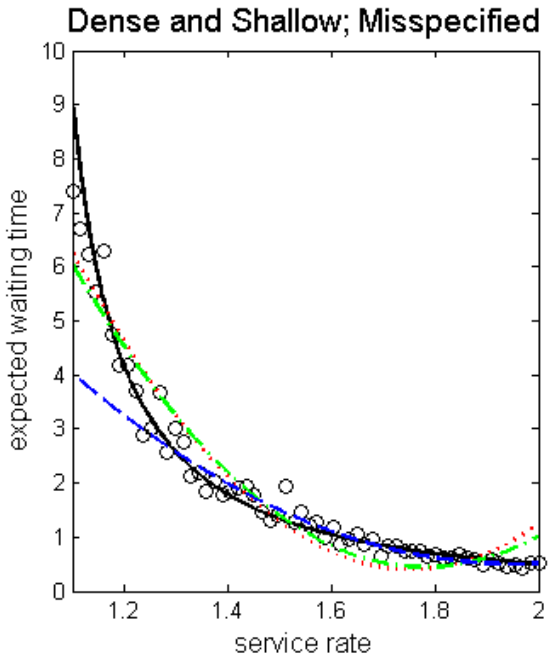
Experiment Design

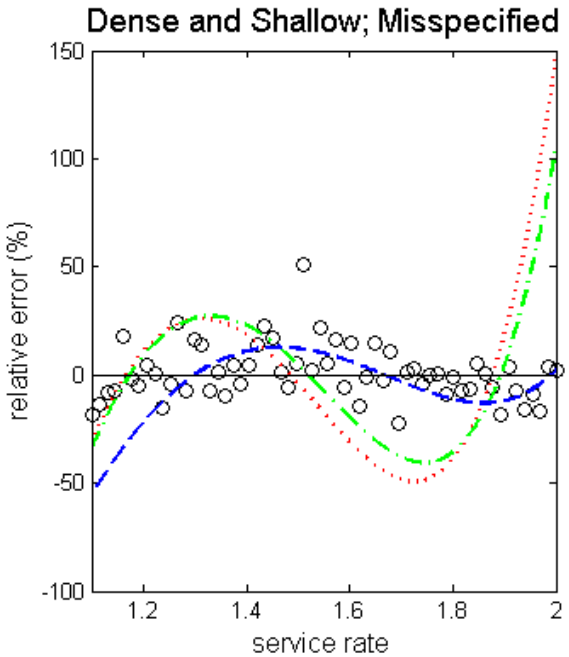
- M/M/1: one-dimensional, evenly spaced
- k design points: x_1=1.1, …, x_k=2
- constant simulation effort n at each
  - 30 replications = runs of n customers each
  - Var[Y(x;n)] is huge for x=1.1, small for x=2
- Two experiment designs:
  - sparse & deep: k=6, n=1000 customers
  - dense & shallow: k=60, n=100 customers

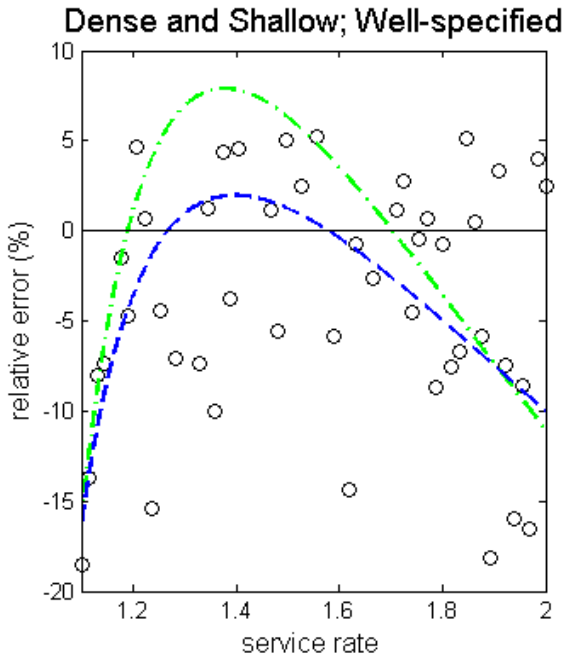

Sparse and Deep; Misspecified
true values
simulation output
LS fit to truth
OLS predictions
EWLS predictions
expected waiting time
service rate


Sparse and Deep; Misspecified
relative error (%)
service rate


Sparse and Deep; Well-specified
relative error (%)
service rate


Dense and Shallow; Misspecified
expected waiting time
service rate


Dense and Shallow; Misspecified
relative error (%)
service rate


Dense and Shallow; Well-specified
relative error (%)
service rate

Regression: Conclusions

- misspecification → poor predictions
  - the WHY of interpolation (including kriging)
- weighted least squares: dangerous!
  - WLS assumes a well-specified model
  - assigns huge residuals to high-variance observations, predicts very badly there
- Want a well-specified model?
  - good luck, hard work

Interpolation

- Deviations from trend are meaningful.
- Can omit trend modeling (overall mean).
- prediction $\hat{y}(x)$ at x after simulating at design points $x_1, \ldots, x_k$:
  - if $x = x_i$ is simulated, $\hat{y}(x_i) = Y(x_i)$
  - if not, $\hat{y}(x) = w_1 Y(x_1) + \ldots + w_k Y(x_k)$ where w_i is larger for x_i closer to x

Kriging: Spatial Corr.

- deterministic simulation:
  - $\varepsilon(x) = Y(x) - y(x) = 0$.
- $Y(\cdot)$ is regarded as a random field
  - each $Y(x)$ is a random variable (Bayesian)
- $\mathrm{Corr}[Y(x), Y(x')] = r(x-x')$
  - e.g. $r(x-x') = \exp(-\sum_i \theta_i |x_i - x'_i|)$
  - or $r(x-x') = \exp(-\sum_i \theta_i (x_i - x'_i)^2)$

Kriging Prediction

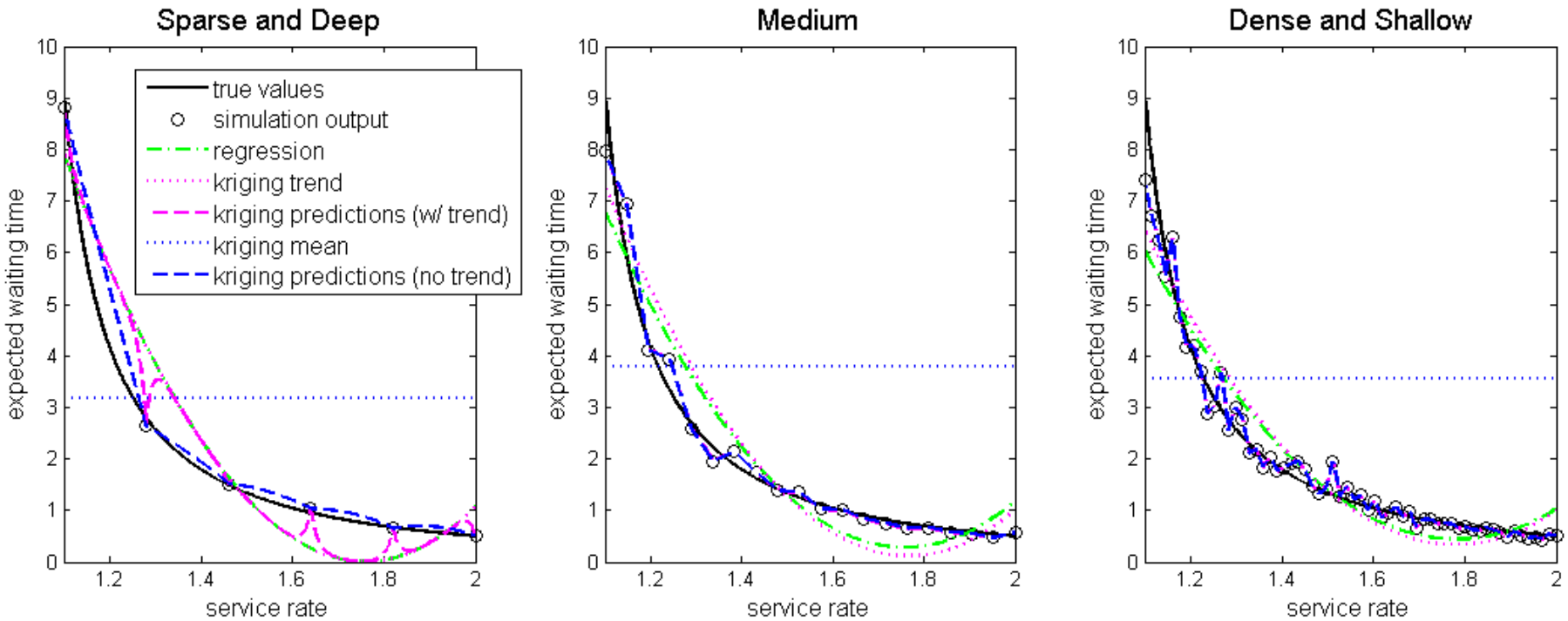
- prior: $Y(\cdot)$ is a Gaussian random field
  - $E[Y(x)] = b(x)\beta$, often $= \beta_0$
  - $Cov[Y(x), Y(x')] = \tau^2 r(x-x')$
- observe $Y = (Y(x_1), \ldots, Y(x_k))$
- posterior mean:
  - $\mathbf{\hat{y}(x) = b(x)\beta + r(x)\, R^{-1}\, (Y - B\beta)}$
  - $r_i(x) = r(x-x_i)$, $R_{ij} = r(x_j-x_i)$, $B_{i\cdot} = b(x_i)$

Choices in Kriging

- basis functions $b(\cdot)$ for trend $b(x)\beta$
  - estimate coefficients β by least-squares: $\min \sum_i (Y(x_i) - b(x)\beta)^2$
- spatial correlation function $r(\cdot;\theta)$
  - estimate coefficients θ by cross-validation or maximizing likelihood of $Y(x_1), \ldots, Y(x_k)$
- axis transformation affects predictions
  - arrival & service rates vs. arrival rate, load

Pathologies of Kriging

- reversion to the trend
- fitting to noise→WHY stochastic kriging



Kriging with Errors

- measurement error or nugget effect
  - $\varepsilon(x) = Y(x) - y(x) \neq 0$
  - filter out noise that harms prediction
  - improve numerical stability
- intrinsic vs. extrinsic uncertainty
  - intrinsic: Var[ε(x)] from physical experiment or noise from stochastic simulation
  - extrinsic: Var[Y(x)] representing our ignorance

How to Apply Kriging to Stochastic Simulation

- Kleijnen & van Beers
  - control noise and its effect on prediction
- Siem & den Hertog
  - reduce sensitivity to stochastic noise
- Yin, Ng & Ng: modified nugget effect
- Ankenman, Nelson & Staum
  - stochastic kriging

Stochastic Kriging

- Intrinsic uncertainty $\varepsilon(x) = Y(x) - y(x)$ is independent of extrinsic uncertainty.
- If simulation effort n at x is large,
  $\varepsilon(x;n) = Y(x;n) - y(x)$
  - is approximately normal
  - we can estimate its variance v(x)/n
  - do empirical weighted least squares

The "What" of SK

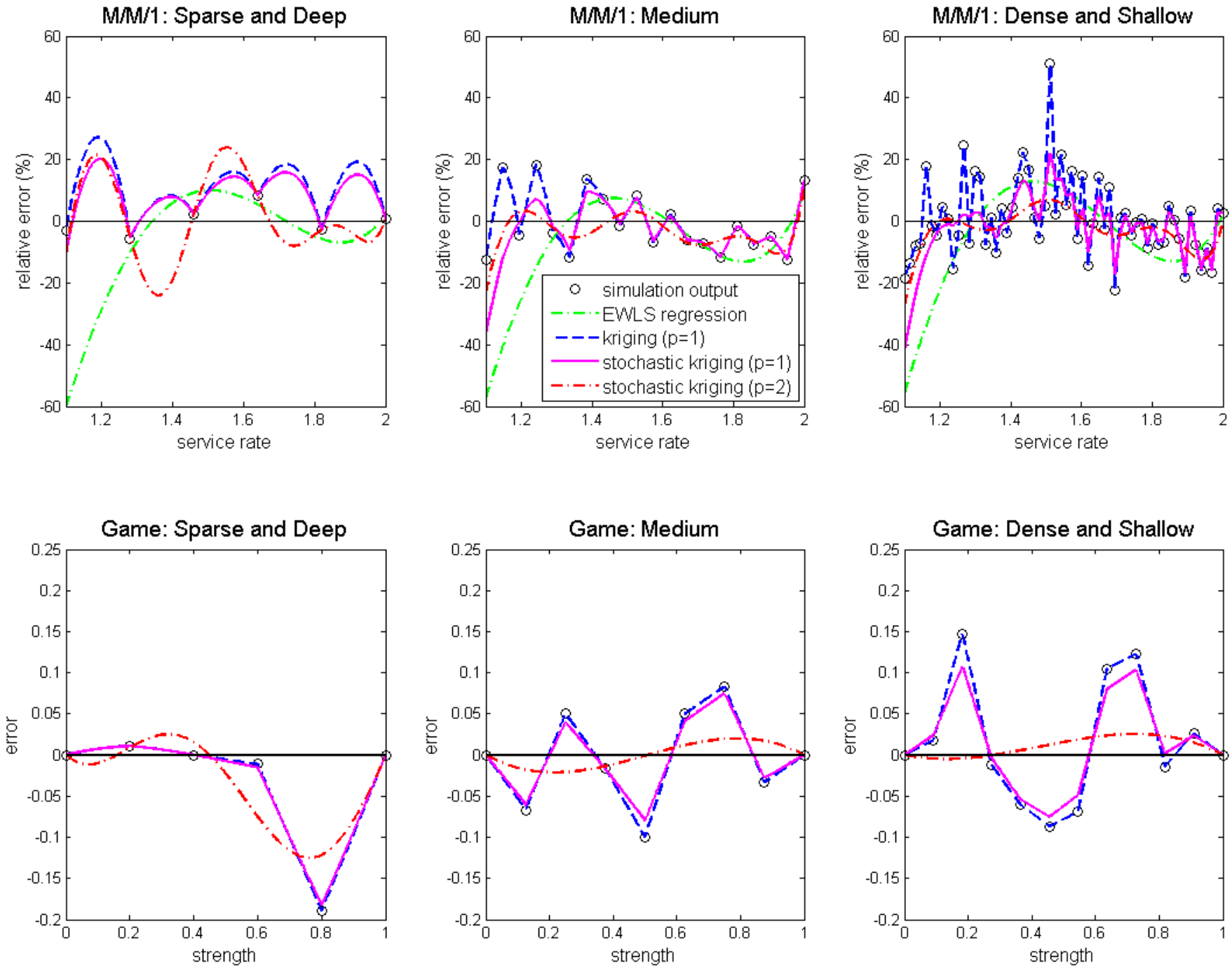
- The kriging prediction
  - $\hat{y}(x) = b(x)\beta + r(x)\, R^{-1} (Y - B\beta)$ where
  - $r_i(x) = r(x-x_i)$, $R_{ij} = r(x_j-x_i)$, $B_{i\cdot} = b(x_i)$
- The stochastic kriging prediction
  - $\hat{y}(x) = b(x)\beta + r(x)\, (R+C/\tau^2)^{-1} (Y - B\beta)$
  - C = intrinsic covariance matrix
  - τ^2 = extrinsic variance
- Awareness of noise alters inference.

Behavior of SK

- If sampling is independent, diagonal C:
  - C_{ii} = intrinsic noise in observing $Y(x_i)$
- The signal-to-noise ratio C_{ii}/τ^2 governs smoothing: how far $\hat{y}(x_i)$ is from $y(x_i)$.
- Extreme cases:
  - $C \downarrow 0$: SK $\rightarrow$ kriging
  - $\tau^2 \downarrow 0$: SK $\rightarrow$ EWLS regression
    - neglecting $Y(x_i)$ if $C_{ii} >> \tau^2$!

Examples

- M/M/1 Queue
  - high intrinsic variance for low service rate x
- a simulated game
  - A tosses a coin, heads with prob. x
  - B tosses another, heads with prob. 1-x
  - HT → B pays A $1; TH → A pays B $1
  - y(x) = 2x-1
  - intrinsic variance v(x) highest in the middle


M/M/1: Sparse and Deep
M/M/1: Medium
M/M/1: Dense and Shallow
relative error (%)
service rate
simulation output
EWLS regression
kriging (p=1)
stochastic kriging (p=1)
stochastic kriging (p=2)
Game: Sparse and Deep
Game: Medium
Game: Dense and Shallow
error
strength

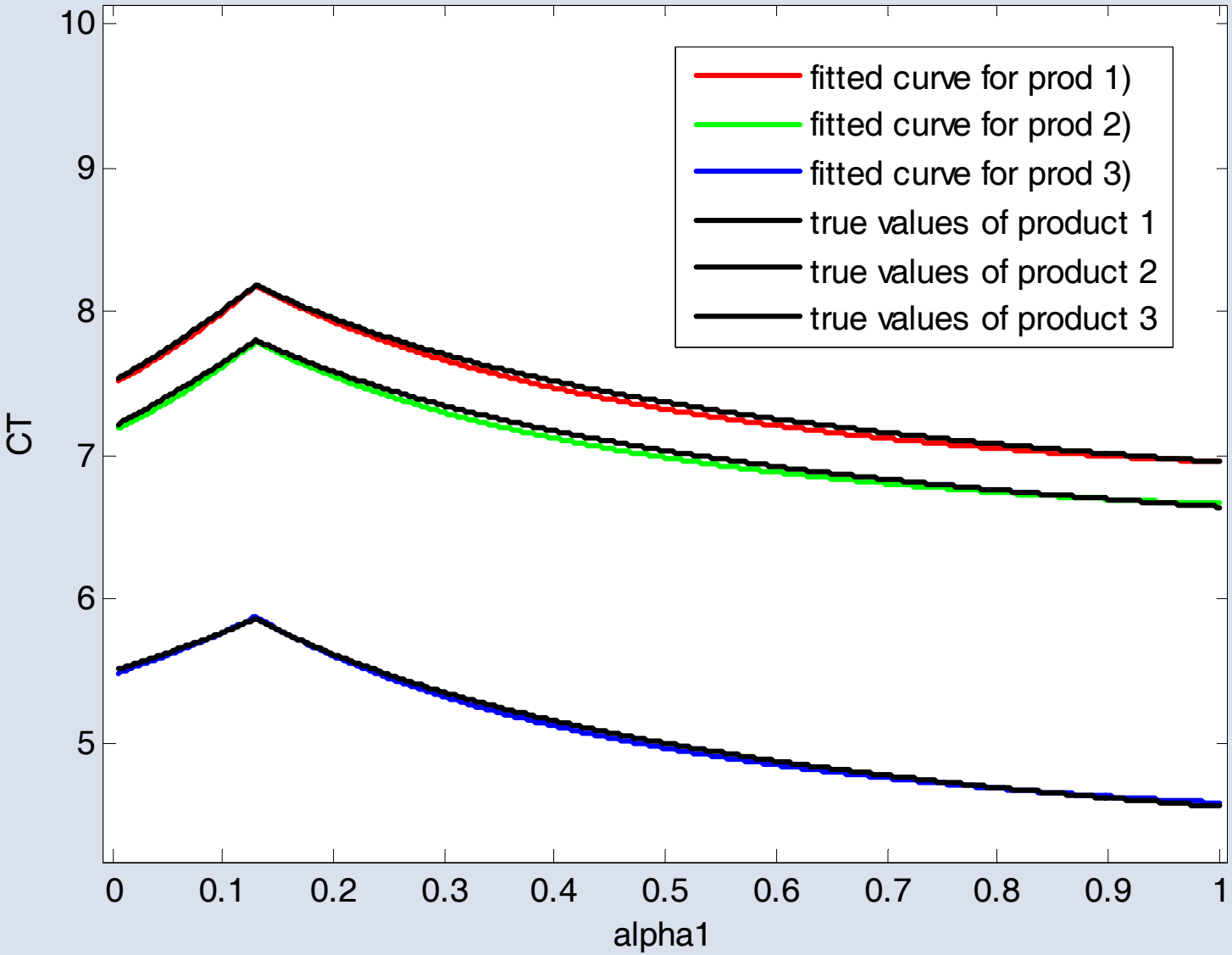
SK: How-To Overview

- pre-simulation:
  - choose axes and correlation function r
  - choose design points and effort: x_i, n_i
- simulate: for each i=1, …, k,
  - observe $Y(x_i)$, estimate variance $v(x_i)$
- post-simulation:
  - estimate parameters β, θ, τ^2
  - predict $\hat{y}(x)$ for any x desired

Multi-Product Cycle Time

- simulate Jackson network with 3 nodes
- input x has 3 dimensions: (did up to 8)
  - mix of 3 products (ranging from 0%-100%)
  - load on bottleneck node (from 0.6-0.8)
- 201 design points (low-discrepancy)
- reduce heteroscedasticity by planning run lengths via approximation
  - intrinsic std. error still varied over 10-fold

α_2:α_3 = 1:6, x = 0.65 (gauss function)



SK: Practical Advice

- Use the Gaussian correlation function
  - $r(x-x') = \exp(-\sum_i \theta_i (x_i - x'_i)^2)$
  - for smoothness of random field, need $r(x-x') \rightarrow 1$ slowly as $x-x' \rightarrow 0$
- misspecification: spatial homogeneity
  - spatial transformation
  - trend modeling?

SK: Practical Advice

- Don't extrapolate! Beware edge effect.
  - bias due to one-sided smoothing near edge
- Placement of design points is important.
  - Grids are not good in high dimension.
  - The goal is not the same as in regression.
  - For uniformity: low-discrepancy sequences.
- Kriging can use lots of CPU time, memory for >1000 design points.

SK: Practical Advice

- Make intrinsic variance $Var[Y(x_i)]/n_i$ smaller than variability of $Y(x_1), \ldots, Y(x_k)$
  - or of $Y(x_1)-b(x_1)\beta, \ldots, Y(x_k)-b(x_k)\beta$
  - Can use two-stage procedure to control intrinsic variance.
- More design points are better than very low variance at a few points.
- Big k, but avoid excessive noise or cost.

SK: Posterior Variance

- Prediction is posterior mean:
  - $\hat{y}(x) = b(x)\beta + r(x)\,(R+C/\tau^2)^{-1}\,(Y - B\beta)$
- Posterior variance:
  - $\tau^2(1 - r(x)\,(R+C/\tau^2)^{-1}\,r(x)^T)$
- Don't trust posterior variance!
  - misleading due to misspecification of GRF
  - can guide effort in multi-stage procedures
  - Kleijnen: bootstrapping & cross-validation

SK: Amelioration

- Spatial inhomogeneity: better-specified
  - data-driven spatial transformation
  - non-stationary covariance functions
  - partitioning: “treed” GRFs
- Computational cost:
  - r(x-x’)=0 if x’ far from x; sparse matrices
  - covariance tapering: set small corr to 0

Conclusions

- Global metamodeling may be feasible.
  - harnessing computer idle time
- Consider stochastic kriging when:
  - regression isn't enough
  - computational budget is limited
  - fewer than 1000 design points (for now)
- www.stochastickriging.net
  - papers, MATLAB code