Introductory Overview Lecture on Computer Experiments -
The Modeling and Analysis of Data from Computer Experiments

Thomas Santner
Ohio State University
Columbus OH

Joint Statistical Meetings
Seattle, WA
August 2006

http://www.stat.ohio-state.edu/~comp_exp
Outline

1. Experiments
2. Experimentation using Computer Codes
3. A Taxonomy of Problems
4. Gaussian Stochastic Process (GaSP) Models
5. Prediction based on the GaSP Model
6. Conclusions-Take Home Messages
1. Experiments

• Physical Experiments
  – **Gold** standard for establishing cause and effect relationships
  – Mainstay of Agriculture, Industry, Medicine
  – Principles of randomization, blocking, choice of sample size, and stochastic modeling of response variables all developed in response to needs of physical experiments
1. Experiments

- **Physical Experiments**
  - *Gold* standard for establishing cause and effect relationships
  - Mainstay of Agriculture, Industry, Medicine
  - Principles of randomization, blocking, choice of sample size, and stochastic modeling of response variables all developed in response to needs of physical experiments

- **Simulation Experiments** Complex physical system each of whose parts interact in a known stochastic manner but whose ensemble is not understood analytically. Heavily used in Industrial Engineering--e.g., compare job shop set ups
1. Experiments

- **Physical Experiments**
  - **Gold** standard for establishing cause and effect relationships
  - Mainstay of Agriculture, Industry, Medicine
  - Principles of randomization, blocking, choice of sample size, and stochastic modeling of response variables all developed in response to needs of physical experiments

- **Simulation Experiments** Complex physical system each of whose parts interact in a known stochastic manner but whose ensemble is not understood analytically. Heavily used in Industrial Engineering--e.g., compare job shop set ups

- **Computer Experiments** relatively new (below)
1. Experiments

• **Physical Experiments**
  – **Gold** standard for establishing cause and effect relationships
  – Mainstay of Agriculture, Industry, Medicine
  – Principles of randomization, blocking, choice of sample size, and stochastic modeling of response variables all developed in response to needs of physical experiments

• **Simulation Experiments** Complex physical system each of whose parts interact in a known stochastic manner but whose ensemble is not understood analytically. Heavily used in Industrial Engineering--e.g., compare job shop set ups

• **Computer Experiments** relatively new (below)

• **Combinations** of the above
2. Experimentation using Computer Codes

• In some situations performing a physical experiment is not feasible
  1. Physical process is technically too difficult to study
  2. Number of variables is too large
  3. Too expensive to study directly (it's all money)
  4. Ethical considerations
2. Experimentation using Computer Codes

• In some situations performing a physical experiment is not feasible
  1. Physical process is technically too difficult to study
  2. Number of variables is too large
  3. Too expensive to study directly (it's all money)
  4. Ethical considerations

• When physical experiments are not possible, it may still be feasible to conduct a **computer experiment**

http://www.stat.ohio-state.edu/~comp_exp
2. Experimentation using Computer Codes

• In some situations performing a physical experiment is not feasible
• When physical experiments are not possible, it may still be feasible to conduct a **computer experiment**

**IF** the physical process relating the inputs \( \mathbf{x} \) to the response(s)

- a. Can be described by a mathematical model relating the output, \( y(\mathbf{x}) \), to \( \mathbf{x} \)
- b. Numerical methods exist for solving the mathematical model
- c. The numerical methods can be implemented with computer code (in finite time!)

**THEN** one can run the computer code to produce a "response" \( y(\mathbf{x}) \) at any input \( \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d \), i.e., one can conduct a computer experiment

\[
\mathbf{x} \longrightarrow \text{Code} \longrightarrow y(\mathbf{x})
\]

The computer code is a **proxy** for the physical process.
Features of Computer Experiments

\[
\begin{align*}
x \longrightarrow & \quad \text{Code} \quad \longrightarrow \quad y(x) \\
\end{align*}
\]

- \( y(x) \) is deterministic
Features of Computer Experiments

\[ x \rightarrow \text{Code} \rightarrow y(x) \]

- \( y(x) \) is deterministic
- \( y(x) \) may be biased (hence the need for "calibration")
Features of Computer Experiments

\[ x \rightarrow \boxed{\text{Code}} \rightarrow y(x) \]

- \( y(x) \) is deterministic
- \( y(x) \) may be biased
- Traditional principles used in designing physical experiments to balance the effects of non-identical experimental units (randomization, blocking, etc) are irrelevant.
Features of Computer Experiments

\[ x \longrightarrow \text{Code} \longrightarrow y(x) \]

- \( y(x) \) is deterministic
- \( y(x) \) may be biased
- Traditional DoE principles are irrelevant
- Sometimes output from a physical experiments is also available. Usual philosophy physical experiment is a noisy measurement of the true input-output relationship \((x \longrightarrow \mu^T(x))\). Model

\[ y^p(x) = \mu^T(x) + \epsilon(x) \quad \text{and} \quad y^c(x) = \mu^T(x) + \delta(x) \]

- \( \{\epsilon(x)\}_x \) are measurement errors (usually, white noise)
- \( \{\delta(x)\}_x \) is computer model bias

Warning
- Sometimes physical experiments are available only for components of the ensemble process, eg, code that emulates an auto crash test.
- In other cases, only experiments that approximate reality are available, e.g., a knee simulator

http://www.stat.ohio-state.edu/~comp_exp
Features of Computer Experiments

\[ x \rightarrow \text{Code} \rightarrow y(x) \]

- \( y(x) \) is deterministic
- \( y(x) \) may be biased
- Traditional DoE principles are irrelevant
- Sometimes output from a physical experiments is also available.
- Our interest in settings where
  - Few computer runs are possible - complex codes (fine-grid FEA codes)
  - High—dimensional input \( x \)
Examples of Computer Experiments

(1) Design of VLSI circuits

(2) Design automobile engines, and other components

Kai-Tai Fang, Runze Li, and Agus Sudjianto (2005) *Design and Modeling for Computer Experiments*

(3) Determine optimum operating conditions for a compression molding process

(4) Determine the performance of controlled nuclear fusion devices

(5) Describe the temporal evolution of contained and wild fires

(6) Design of jet engines, helicopter rotor blades

(7) Biomechanics — Explain behavior of (or even design) prosthetic devices

http://www.stat.ohio-state.edu/~comp_exp
Example  Zone Computer Models used to predict the evolution of a fire in an enclosed room (eg, the computer code ASET = Available Safe Egress Time)

In particular, ASET-B describes the temporal evolution of a fire in a single room with closed doors and windows that contains an object at some point below the ceiling that has been ignited.

Fig. 3-10.1. Events immediately after ignition.
Fig. 3-10.3.  The plume-ceiling interaction.
Fig. 3-10.4. Ceiling jet-wall interaction.
Fig. 3-10.6. Further "smoke filling."
Mathematical Model

Inputs to ASET-B

• Room ceiling height \(x_1\)
• Room floor area \(x_2\)
• Height of the fire source (the burning object) above the floor \(x_3\)
• Heat loss fraction for the room (depends on room insulation) \(x_4\)
• (etc, material-specific heat release rate)

Computed Response

\[
y(x_1, x_2, x_3, x_4) = \text{time required by smoke layer to reach 5 ft above the ground}
\]
**Objective**  Predict the time for the fire plume to reach 5 ft above the ground for untried combinations

http://www.stat.ohio-state.edu/~comp_exp
3. A Taxonomy of Problems

Setup

1. Inputs $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_e, \mathbf{x}_m)$ where
   - $\mathbf{x}_c =$ control (manufacturing, engineering design) variables
   - $\mathbf{x}_e =$ noise (field, environmental) variables
   - $\mathbf{x}_m =$ model variables

   *(Not all types of inputs need be present in every application.)*
3. A Taxonomy of Problems

Setup

1. **Inputs** \( \mathbf{x} = (x_c, x_e, x_m) \) where
   \( x_c \) = control (manufacturing, engineering design) variables
   \( x_e \) = noise (field, enviromental) variables
   \( x_m \) = model and calibration variables
   *(Not all types of inputs need be present in every application.)*

2. **Outputs**
   - Real-valued : \( y(\mathbf{x}) \) *or*
   - Multivariate: \( (y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_k(\mathbf{x})) \) *or*
   - Functional: \( (t, y(t, \mathbf{x})) \)
3. A Taxonomy of Problems

Setup

1. Inputs $x = (x_c, x_e, x_m)$ where
   - $x_c =$ control (manufacturing, engineering design) variables
   - $x_e =$ noise (field, environmental) variables
   - $x_m =$ model variables

(Not all types of inputs need be present in every application.)

2. Outputs
   - Real-valued: $y(x)$ or
   - Multivariate: $(y_1(x), y_2(x), \ldots, y_k(x))$ or
   - Functional: $(t, y(t, x))$

3. $X_e \sim F(\cdot)$ describes "target field conditions"
   $X_m \sim \pi(\cdot)$ describes prior knowledge about the model variables
3. A Taxonomy of Problems

Setup

1. Inputs \( \mathbf{x} = (\mathbf{x}_c, \mathbf{x}_e, \mathbf{x}_m) \) where
   - \( \mathbf{x}_c = \text{control (manufacturing, engineering design)} \) variables
   - \( \mathbf{x}_e = \text{noise (field, environmental)} \) variables
   - \( \mathbf{x}_m = \text{model} \) variables

(Not all types of inputs need be present in every application.)

2. Outputs
   - Real-valued: \( y(\mathbf{x}) \) or
   - Multivariate: \( (y_1(\mathbf{x}), y_2(\mathbf{x}), \ldots, y_k(\mathbf{x})) \) or
   - Functional: \( (t, y(t, \mathbf{x})) \)

3. \( \mathbf{X}_e \sim F(\bullet) \) describes "target field conditions"
   \( \mathbf{X}_m \sim \pi(\bullet) \) describes prior knowledge about the model specifications

4. Summary of the \( y(\mathbf{x}_c, \mathbf{X}_e) \) distribution (assuming no \( \mathbf{X}_m \))
   \[
   \begin{align*}
   \mu(\mathbf{x}_c) &= E_F\{y(\mathbf{x}_c, \mathbf{X}_e)\}; \\
   \xi(\mathbf{x}_c) : P_F\{y(\mathbf{x}_c, \mathbf{X}_e) \leq \xi(\mathbf{x}_c)\} &= \alpha \text{ (median)}; \\
   \sigma^2(\mathbf{x}_c) &= \text{Var}_F(y(\mathbf{x}_c, \mathbf{X}_e))
   \end{align*}
   \]
3. A Taxonomy of Problems

Setup

1. **Inputs** \( x = (x_c, x_e, x_m) \) where
   - \( x_c \) = control (manufacturing, engineering design) variables
   - \( x_e \) = noise (field, environmental) variables
   - \( x_m = (x_{m,\text{math}}, x_{m,\text{num}}) \) = model variables
(Not all types of inputs need be present in every application.)

2. **Outputs**
   - Real-valued: \( y(x) \) *or*
   - Multivariate: \( (y_1(x), y_2(x), \ldots, y_k(x)) \) *or*
   - Functional: \( (t, y(t, x)) \)

3. \( X_e \sim F(\cdot) \) describes "target field conditions"
   \( X_m \sim \pi(\cdot) \) describes prior knowledge about the model specifications

4. Summary (assuming no \( X_m \)) of \( y(x_c, X_e) \) distribution
   \( \mu(x_c) = E_F\{y(x_c, X_e)\} \);
   \( \xi(x_c) : P_F\{y(x_c, X_e) \leq \xi(x_c)\} = \alpha \) (median);
   \( \sigma^2(x_c) = \text{Var}_F(y(x_c, X_e)) \)

5. "Design" of a computer experiment \( \equiv \) choice of \( x^t_1, \ldots, x^t_n \) at which to
    evaluate computer code, where, eg, \( x^t_i = (x^t_{c,i}, x^t_{e,i}) \), \( i = 1, \ldots, n \)
3. Taxonomy of Problems

Problem 1 Interpolation/Prediction — Given computer code output at a set of training inputs,

\[(x_1^t, y(x_1^t)), \ldots, (x_n^t, y(x_n^t))\]

predict \(y(\cdot)\) at a new input \(x_0\) (predictor \(\equiv \text{metamodel}\))
3. Taxonomy of Problems

Problem 1 Interpolation/Prediction — Given computer code output at a set of "training" inputs,

\[(x_1^t, y(x_1^t)), \ldots, (x_n^t, y(x_n^t))\]

predict \(y(\bullet)\) at a new input \(x_0\) (predictor \(\equiv\) metamodel)

Problem 2 Experimental design — Determine a set of inputs at which to carry out the sequence of code runs (a "good" design of a physical or computer experiment depends on the scientific objective of the research)

— Exploratory Designs ("space-filling")
— Prediction-based Designs
— Optimization-based Designs (e.g., find \(x_{opt}^c = \text{argmin } y(x)\))
3. Taxonomy of Problems

Problem 1  Interpolation/Prediction

Problem 2  Experimental design
  – Exploratory Designs ("space-filling")
  – Prediction-based Designs
  – Optimization-based Designs (e.g., find $x_c^{opt} = \arg\min y(x)$)

Problem 3  Uncertainty/Output Analysis — Determine the distribution of
the random variable $y(x_c, X_e)$. (Determine the variability in the performance
measure $y(\bullet)$ for design $x_c$ when applied to the population defined by the
distribution of $X_e$, eg, patient specific variables (patient weight or bone
material properties) or surgeon specific variables (measuring surgical skill)
Example In his Cornell PhD thesis, Kevin Ong conducted an uncertainty analysis of the effect of Engineering Cup design, Surgical, Patient, and Fluid Effects on the Stability of Uncemented Acetabular Components.
3. Taxonomy of Problems

Problem 1 Interpolation/Prediction

Problem 2 Experimental design

Problem 3 Uncertainty/Output Analysis

Problem 4 Sensitivity Analysis — Determine how variation in $y(x)$ can be apportioned to the different inputs of $x$ (which inputs is $y(x)$ not sensitive to? which sets of inputs is $y(x)$ most sensitive to?)

Philosophy Inputs that have relatively little effect on the output can be set to some nominal value; additional investigation can be restricted to determining how the output depends on the active inputs
• $r$ = Room ceiling height
• $a$ = Room floor area
• $f$ = Height of the fire source above the floor
• $h$ = Heat loss fraction for the room

http://www.stat.ohio-state.edu/~comp_exp
3. Taxonomy of Problems

Problem 1 Interpolation/Prediction

Problem 2 Experimental design

Problem 3 Uncertainty/Output Analysis

Problem 4 Sensitivity Analysis

Problem 5 Calibrate the computer code — Use outputs from both a physical experiment and an associated computer code that represents the physical process to set the computer code calibration variables. Calibration variables are those $x_m$ variables that can be used to eliminate the bias in a computer code's representation of a physical input-output relationship.

Typical Numerical Variables

1. FEA Mesh Density = ?
2. Load Discretization = ?
3. Solution tolerances = ?, etc
3. Taxonomy of Problems

Problem 1 Interpolation/Prediction

Problem 2 Experimental design

Problem 3 Uncertainty/Output Analysis

Problem 4 Sensitivity Analysis

Problem 5 Calibrate the computer code

Problem 6 Prediction Accuracy – Using the data from both a physical experimental data and a (calibrated) computer experiment, give predictions (including uncertainty bounds) for an associated physical system.

http://www.stat.ohio-state.edu/~comp_exp
3. Taxonomy of Problems

Problem 1 Interpolation/Prediction

Problem 2 Experimental design

Problem 3 Uncertainty/Output Analysis

Problem 4 Sensitivity Analysis

Problem 5 Calibration

Problem 6 Prediction

Problem 7 Find Robust Inputs — In experiments with engineering design and patient-specific environmental variables, determine robust choices of the $x_c$ engineering design variables. If

$$\mu(x_c) = E_F\{y(x_c, X_e)\}$$

then a robust set of inputs $x_c$ is an engineering "design" whose output is minimally sensitive to the assumed distribution $F(\cdot)$ of $X_e$
**Bottom Line** Many of the problems above have "natural" solutions obtained by approximating $y(\mathbf{x}_c, \mathbf{x}_e)$ by a fast (linear in the training data) predictor, a metamodel.
4. Gaussian Stochastic Process (GaSP) Models  
(used as basis for both prediction and some design choices)

Idea  Regard $y(\mathbf{x})$ as a realization, a "draw," of a random function $Y(\mathbf{x})$  
(Bayesian Thinking)

The simplest possible (prior) model for $Y(\mathbf{x})$ is

$$Y(\mathbf{x}) = \sum_j \beta_j f_j(\mathbf{x}) + Z(\mathbf{x})$$

"large scale trends"  
"smooth deviations"

$$= \beta^T f(\mathbf{x}) + Z(\mathbf{x})$$

where

$f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ are known regression functions, 
$\boldsymbol{\beta}$ is an unknown regression vector, and 
$Z(\mathbf{x})$ is a stationary Gaussian Stochastic Process (GaSP)
\[ Z(\mathbf{x}), \mathbf{x} \in \mathcal{X} \text{ satisfies} \]
\[ \circ \ E\{Z(\mathbf{x})\} = 0 \text{ (zero mean)} \]
\[ \quad (\Rightarrow E\{Y(\mathbf{x})\} = \mathbf{\beta}^\top f(\mathbf{x}) + 0 = \mathbf{\beta}^\top f(\mathbf{x})) \]
\[ \circ \ \text{Var}(Z(\mathbf{x})) = \sigma^2_Z \]
\[ \circ \ \text{Correlation Function: symmetric } R(\cdot) \text{ with } R(0) = 1, \]
\[ \quad \text{Cov}(Z(\mathbf{x}_1),Z(\mathbf{x}_2)) = \sigma^2_Z \times R(\mathbf{x}_1 - \mathbf{x}_2) \]
\[ \quad (\sigma^2_Z = \text{Var}(Z(\mathbf{x}))) \]
\[ \circ \ \text{Typically } R(\cdot) = R(\cdot | \xi) \text{ is a function of a finite number of unknown parameters} \]
\[ \circ \ \text{GaSP: For any } \mathbf{x}_1, \ldots, \mathbf{x}_s, (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_s)) \text{ has the multivariate normal distribution} \]

\[ \bullet \ \text{Usually, taking } \mathbf{\beta}^\top f(\mathbf{x}) = \beta_0 \text{ with a data-selected parametric correlation function } R(\cdot | \xi) \]
4. GaSP Models

**GaSP Models Are Flexible** Four draws from $Z(x)$, a zero mean, unit variance GaSP with input $x \in [0,1]$ and having correlation function

$$R(h) = \exp(-\theta h^2)$$

for $\theta \in \{4, 16, 100\}$
• Some draws from a GaSP with inputs $\mathbf{x} \in [0, 1]^2$
4. GaSP Models

- GaSP Models Are Flexible

- GaSP Models are computationally "simple"
4. GaSP Models

• GaSP Models Are Flexible

• GaSP Models are computationally "simple"

• GaSP Models are Bayesian in character
5. Prediction based on the GaSP Model

- Given (training) data

\[(x_1^t, y(x_1^t)), \ldots, (x_n^t, y(x_n^t))\]

predict \(y(x_0)\), where \(x_0\) is an untried new input.
5. Prediction based on the GaSP Model

- Given (training) data

\[(x_1^t, y(x_1^t)), \ldots, (x_n^t, y(x_n^t))\]

predict \(y(x_0)\), where \(x_0\) is an untried new input.

- Notation \(Y^n = (Y(x_1^t), \ldots, Y(x_n^t))\) and \(y^n = (y(x_1^t), \ldots, y(x_n^t))\)
5. Prediction based on the GaSP Model

- Given (training) data

\[(\mathbf{x}_1^t, y(\mathbf{x}_1^t)), \ldots, (\mathbf{x}_n^t, y(\mathbf{x}_n^t))\]

predict \(y(\mathbf{x}_0)\), where \(\mathbf{x}_0\) is an untried new input.

- Notation \(\mathbf{Y}^n = (Y(\mathbf{x}_1^t), \ldots, Y(\mathbf{x}_n^t))\) and \(\mathbf{y}^n = (y(\mathbf{x}_1^t), \ldots, y(\mathbf{x}_n^t))\)

- The minimum MSPE predictor of \(y(\mathbf{x}_0)\) is

\[\hat{y}(\mathbf{x}_0) = E\{Y(\mathbf{x}_0) \mid \mathbf{Y}^n = \mathbf{y}^n\}\]
5. Prediction based on the GaSP Model

• Given (training) data

\[
(x_1^t, y(x_1^t)), \ldots, (x_n^t, y(x_n^t))
\]

predict \( y(x_0) \), where \( x_0 \) is an untried new input.

• Notation \( Y^n = (Y(x_1^t), \ldots, Y(x_n^t)) \) and \( y^n = (y(x_1^t), \ldots, y(x_n^t)) \)

• The minimum MSPE predictor of \( y(x_0) \) is

\[
\hat{y}(x_0) = E\{Y(x_0) \mid Y^n = y^n\}
\]

Example If \( Y(x) \) follows the GaSP(\( \beta_0, \sigma^2_z, R(\cdot) \)), then

\[
\hat{y}(x_0) \equiv E\{Y(x_0) \mid Y^n = y^n\} = \hat{\beta}_0 + r^\top(x_0)R^{-1}\left(y^n - \hat{\beta}_0 1_n\right)
\]

where

• \( R = (R(x_i^t - x_j^t)) \) is \( n \times n \)
• \( r^\top(x_0) = (R(x_0 - x_1^t), \ldots, R(x_0 - x_n^t)) \) is \( 1 \times n \)
• \( \hat{\beta}_0 \equiv \text{WLSE of } \beta_0 = (1_n^\top R^{-1} 1_n)^{-1}(1_n^\top R^{-1} y^n) \)

http://www.stat.ohio-state.edu/~comp_exp
5. Prediction based on the GaSP Model

- Given \textbf{(training)} data
  \[ (x_1^t, y(x_1^t)), \ldots, (x_n^t, y(x_n^t)) \]
  
  predict \( y(x_0) \), where \( x_0 \) is an untried new input.

- Notation \( Y^n = (Y(x_1^t), \ldots, Y(x_n^t)) \) and \( y^n = (y(x_1^t), \ldots, y(x_n^t)) \)

- The minimum MSPE predictor of \( y(x_0) \) is
  \[ \hat{y}(x_0) = E\{Y(x_0) \mid Y^n = y^n\} \]

**Example** If \( Y(x) \) follows the GaSP(\( \beta_0, \sigma_z^2, R(\cdot) \)), then

\[ \hat{y}(x_0) \equiv E\{Y(x_0) \mid Y^n = y^n\} = \hat{\beta}_0 + r(x_0)R^{-1}(y^n - \hat{\beta}_0 1_n) \]

- **IF** only the \textbf{moment assumptions} holds,
  \( \hat{y}(x_0) \equiv \text{Best Linear Unbiased Predictor} \) of \( Y(x_0) \)
5. Prediction based on the GaSP Model

- **Model Prediction uncertainty** at $\mathbf{x}_0$

\[
\sigma^2(\mathbf{x}_0) = \mathbb{E}\{(Y(\mathbf{x}_0) - \hat{y}(\mathbf{x}_0))^2 \mid \mathbf{Y}^n = \mathbf{y}^n\}
\]
5. Prediction based on the GaSP Model

• Model Prediction uncertainty at $\mathbf{x}_0$

$$\sigma^2(\mathbf{x}_0) = \mathbb{E}\{(Y(\mathbf{x}_0) - \widehat{y}(\mathbf{x}_0))^2 \mid Y^n = y^n\}$$

• Empirical BLUP If the correlation is unknown (the usual case),

$\Rightarrow \mathbf{R}, \mathbf{r}^\top(\mathbf{x}_0)$, and $\widehat{\beta}_0$ are also unknown:

If, further, $R(\cdot) = R(\cdot \mid \xi)$ is parametric, and we estimate $\xi$ by $\widehat{\xi}$, say, we can predict using the corresponding empirical BLUP

$$\widehat{y}(\mathbf{x}_0) \equiv \mathbb{E}\{Y(\mathbf{x}_0) \mid Y^n = y^n, \widehat{\xi}\} = \widehat{\beta}_0 + \mathbf{r}^\top(\mathbf{x}_0)\widehat{\mathbf{R}}^{-1}\left(y^n - \widehat{\beta}_0 \mathbf{1}_n\right)$$

$\widehat{\xi} = \text{MLE, REML, penalized likelihood, or other estimator of } \xi$
5. Prediction based on the GaSP Model

- **Model Prediction uncertainty** at $x_0$

\[ \sigma^2(x_0) = \mathbb{E}\{(Y(x_0) - \hat{y}(x_0))^2 \mid Y^n = y^n\} \]

- **Empirical BLUP** If the correlation is unknown (the usual case),

\[ \Rightarrow R, r^T(x_0), \text{ and } \hat{\beta}_0 \text{ are also unknown : } - ( \]

If, further, $R(\bullet) = R(\bullet \mid \xi)$ is parametric, and we estimate $\xi$ by $\hat{\xi}$, say, we can predict using the corresponding empirical BLUP

\[ \hat{y}(x_0) \equiv \mathbb{E}\{Y(x_0) \mid Y^n = y^n, \hat{\xi}\} = \hat{\beta}_0 + \hat{r}(x_0)\hat{R}^{-1}\left(y^n - \hat{\beta}_0 1_n\right) \]

$\hat{\xi} = \text{MLE, REML, penalized likelihood, or other estimator of } \xi$

- **Fully Bayesian Predictor** (giving all parameters priors)

\[ \hat{y}(x_0) = \mathbb{E}\{Y(x_0) \mid Y^n = y^n\} = \mathbb{E}\left\{ \mathbb{E}\{Y(x_0) \mid \beta_0, \sigma^2_z, \xi, y^n\} \mid \beta_0, \sigma^2_z, \xi, y^n\right\}\]

Need $[\beta_0, \sigma^2_z, \xi \mid y^n]$
Properties of $\hat{y}(x_0) = \hat{\beta}_0 + r^\top(x_0)R^{-1}(y^n - \hat{\beta}_0 1_n)$

- Simple to compute (linear in $y^n$)

$$\hat{y}(x) = c_0(x) + \sum_{j=1}^{n} c_j(x) y(x^t_j)$$
Properties of $\widehat{y}(x_0) = \widehat{\beta}_0 + r^\top(x_0)R^{-1}(y^n - \widehat{\beta}_0 1_n)$

- Simple to compute (linear in $y^n$)

$$\widehat{y}(x_0) = c_0(x_0) + \sum_{j=1}^{n} c_j(x_0) y(x_j^t)$$

- but, unfortunately, not the Empirical BLUP :-(
- GASP (W. Welch)
  - PErK (B. J. Williams)
  - SAS Proc Mixed
  - BACCO (Hankin)
  - and others........
- $R^{-1}$ can be computationally demanding
Properties of \( \hat{y}(x_0) = \hat{\beta}_0 + r^\top(x_0)R^{-1}(y^n - \hat{\beta}_0 1_n) \)

- Simple to compute (linear in \( y^n \))

\[
\hat{y}(x_0) = c_0(x_0) + \sum_{j=1}^{n} c_j(x_0) y(x_j^t)
\]

- Viewed as a function of \( x_0 \),

\[
\hat{y}(x_0) = d_0 + \sum_{j=1}^{n} d_j R(x_0 - x_j^t)
\]
Properties of $\hat{y}(x_0) = \hat{\beta}_0 + r^\top(x_0)R^{-1}(y^n - \hat{\beta}_0 1_n)$

• Simple to compute (linear in $y^n$)

$$\hat{y}(x_0) = c_0(x_0) + \sum_{j=1}^{n} c_j(x_0) y(x_j^t)$$

• Viewed as a function of $x$,

$$\hat{y}(x_0) = d_0 + \sum_{j=1}^{n} d_j R(x_0 - x_j^t)$$

• $\hat{y}(x)$ interpolates data, i.e.,

$$\hat{y}(x_j^t) = y(x_j^t), \; j = 1, \ldots, n$$
Properties of \( \hat{y}(x_0) = \beta_0 + r^\top(x_0)R^{-1}(y^n - \beta_0 1_n) \)

• Simple to compute (linear in \( y^n \))

\[
\hat{y}(x_0) = c_0(x_0) + \sum_{j=1}^n c_j(x_0)y(x_j^t)
\]

• Viewed as a function of \( x \),

\[
\hat{y}(x_0) = d_0 + \sum_{j=1}^n d_j R(x_0 - x_j^t)
\]

• \( \hat{y}(x) \) interpolates data, i.e.,

\[
\hat{y}(x_j^t) = y(x_j^t), \ j = 1, \ldots, n
\]

• Splines, neural networks and other well-known interpolators correspond to specific choices of regressors and correlation function \( R(\bullet) \)
Illustration True Curve (solid); five training data points (diamonds); REML-EBLUP with exponential correlation function (dotted)
5. Conclusions-Take Home Messages

1. An increasing number of phenomenon that could previously be studied only by physical experiments, can now be investigated using "computer experiments" or combinations of computer and physical experiments.

2. Modeling the responses from computer experiments must account for the (highly) correlated nature of the output \( y(\mathbf{x}) \) over the input space.

3. Prediction of the output function \( y(\mathbf{x}) \) based on Gaussian (or other) stochastic processes can be used to *interpolate* training data.

4. GaSP models are the basis for
   - Determining interpolating predictors of \( y(\mathbf{x}) \)
   - Assessing error bands of the predictor due to model uncertainty,
   - Targeted experimental design, and
   - Solving calibration problems

http://www.stat.ohio-state.edu/~comp_exp
Related Sessions

335 Multivariate Functional ANOVA for Kriging Model in Computer Experiments (Zhe Zhang)

345 Statistical Design of Computer Experiments for a 3D Chemical Microanalysis Imaging System (Juan Soto, James J. Filliben, and John H. Scott)

363 Design and Analysis of Experiments for Complex Computer Simulators (Aug 9 Session)
   • Designs for Integrated Computer and Physical Experiments (C. Shane Reese, Derek Bingham, and Wilson Lu)
   • Sequential Experiment Design for Contour Estimation from Complex Computer Codes (Pritam Ranjan)
   • Uncertainty Quantification for Combining Experimental Data and Computer Simulations from Multiple Data Sources (Brian J. Williams, Dave Higdon, Jim Gattiker)

474 Network Analysis and Spatial Applications-Contributed (Aug 9 Session)
   • Calibration and Prediction for Computer Experiment Output Having Qualitative and Quantitative Input Variables (Gang Han, Thomas Santner, and William Notz)
   • Validity of Likelihood and Bayesian Inference for Gaussian Process Regression (Bela Nagy, Jason Loeppky, and William J. Welch)

Questions?

http://www.stat.ohio-state.edu/~comp_exp