Knowledge Discovery from Simulations

Donald E. Brown
Predictive Technology Laboratory
University of Virginia
Charlottesville, VA 22904
brown@virginia.edu

October 12, 2009
<table>
<thead>
<tr>
<th>Agenda for the Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiments and Simulations</td>
</tr>
<tr>
<td>Models for Simulation Output</td>
</tr>
<tr>
<td>Design of Simulation Experiments</td>
</tr>
<tr>
<td>Approaches to Prediction from Simulations</td>
</tr>
<tr>
<td>Example Learning from Simulations</td>
</tr>
</tbody>
</table>
Physical Experiments

- A prospective designed experiment is the scientifically accepted method for establishing cause and effect relationships.
Physical Experiments

- A prospective designed experiment is the scientifically accepted method for establishing cause and effect relationships.
- Agriculture provided many of the first uses of designed experiments to understand and change (improve?) practice.
Physical Experiments

- A prospective designed experiment is the scientifically accepted method for establishing cause and effect relationships.
- Agriculture provided many of the first uses of designed experiments to understand and change (improve?) practice.
- Manufacturing and industrial processes used designed experiments to increase productivity and later quality.
Physical Experiments

- A prospective designed experiment is the scientifically accepted method for establishing cause and effect relationships.
- Agriculture provided many of the first uses of designed experiments to understand and change (improve?) practice.
- Manufacturing and industrial processes used designed experiments to increase productivity and later quality.
- In the 1950’s health care adopted clinical trials to understand the effects of treatments.
Physical Experiments

- A prospective designed experiment is the scientifically accepted method for establishing cause and effect relationships.
- Agriculture provided many of the first uses of designed experiments to understand and change (improve?) practice.
- Manufacturing and industrial processes used designed experiments to increase productivity and later quality.
- In the 1950’s health care adopted clinical trials to understand the effects of treatments.
- Education, transportation, and others have employed similar methods, although in limited contexts.
Key Components for Physical Experiments

- **Randomization**: prevent nuisance variables from confounding the impact of the control or treatment variables on the response.
Key Components for Physical Experiments

- **Randomization**: prevent nuisance variables from confounding the impact of the control or treatment variables on the response.
- **Blocking**: directly controls the effects of recognized nuisance variables on the response.
Key Components for Physical Experiments

- **Randomization**: prevent nuisance variables from confounding the impact of the control or treatment variables on the response.
- **Blocking**: directly controls the effects of recognized nuisance variables on the response.
- **Replication**: reduces the possible obscuration of the relationship between the control variables and the response by variation in the measurement of the response.
Simulations and Computer Experiments

- Originally computer experiments provided a means for abstractly relating the output from a complex physical process to a set of input variables (e.g., systems of differential equations).
## Simulations and Computer Experiments

- Originally computer experiments provided a means for abstractly relating the output from a complex physical process to a set of input variables (e.g., systems of differential equations).
- Why simulate?
Simulations and Computer Experiments

- Originally computer experiments provided a means for abstractly relating the output from a complex physical process to a set of input variables (e.g., systems of differential equations).
- Why simulate?
  - We want to find relationships between input and output variables in complex problems. This will become clear as we look at some examples.
Simulations and Computer Experiments

- Originally computer experiments provided a means for abstractly relating the output from a complex physical process to a set of input variables (e.g., systems of differential equations).

- Why simulate?
  - We want to find relationships between input and output variables in complex problems. This will become clear as we look at some examples.
  - The need for quick understanding of complex problems, and sometimes even simple problems, motivates the use of simulations. Consider the simple Monty Hall problem.
Suppose you’re a contestant on Let’s Make A Deal, and Monty Hall gives you the choice of three doors: Behind one door is the grand prize: a banquet for you and your family; behind the other two doors are goats.
Monty Hall Problem

• Suppose you’re a contestant on Let’s Make A Deal, and Monty Hall gives you the choice of three doors: Behind one door is the grand prize: a banquet for you and your family; behind the other two doors are goats.
• You pick a door. Then Monty opens one of the remaining doors to reveal a goat.
Suppose you’re a contestant on Let’s Make A Deal, and Monty Hall gives you the choice of three doors: Behind one door is the grand prize: a banquet for you and your family; behind the other two doors are goats.

You pick a door. Then Monty opens one of the remaining doors to reveal a goat.

Should you keep the door you selected or change?
Bayesian Solution

Suppose you choose door A and learn that a goat is behind door C. Should you change doors?
Bayesian Solution

Suppose you choose door A and learn that a goat is behind door C. Should you change doors?

\[
P(A|C^c) = \frac{P(A)P(C^c|A)}{P(C^c)} = \frac{(1/3) \times 1}{2/3} = 0.5
\]

So it does not matter.
Bayesian Solution

Suppose you choose door A and learn that a goat is behind door C. Should you change doors?

\[
P(A|C^c) = \frac{P(A)P(C^c|A)}{P(C^c)} = \frac{(1/3) \times 1}{(2/3)} = .5
\]

So it does not matter. But this is wrong!
Set WinStay and WinSwitch = 0.
Repeat 10,000 times
Draw u \sim U(0, 1)
If u > 0.33
Then WinSwitch ← WinSwitch + 1
Else WinStay ← WinStay + 1
WinStay ← WinStay/10000;
WinSwitch ← WinSwitch/10000
Simulation Solution

Set WinStay and WinSwitch = 0.

Repeat 10,000 times
Draw $u \sim U(0, 1)$
If $u > 0.33$
Then WinSwitch ← WinSwitch + 1
Else WinStay ← WinStay + 1
WinStay ← WinStay/10000;
WinSwitch ← WinSwitch/10000

- Running this simulation, for instance, gives WinStay = .3328 and WinSwitch = .6672.
Simulation Solution

Set $\text{WinStay}$ and $\text{WinSwitch} = 0$.
Repeat 10,000 times
Draw $u \sim U(0, 1)$
If $u > 0.33$
Then $\text{WinSwitch} \leftarrow \text{WinSwitch} + 1$
Else $\text{WinStay} \leftarrow \text{WinStay} + 1$
$\text{WinStay} \leftarrow \text{WinStay}/10000$;
$\text{WinSwitch} \leftarrow \text{WinSwitch}/10000$

- Running this simulation, for instance, gives $\text{WinStay} = 0.3328$ and $\text{WinSwitch} = 0.6672$.
- The chance that the door you first picked is a winner is $1/3$. 

$\text{Set } \text{WinStay} \text{ and } \text{WinSwitch} = 0$. 
\text{Repeat 10,000 times} 
\text{Draw } u \sim U(0, 1) 
\text{If } u > 0.33 
\text{Then } \text{WinSwitch} \leftarrow \text{WinSwitch} + 1 
\text{Else } \text{WinStay} \leftarrow \text{WinStay} + 1 
\text{WinStay} \leftarrow \text{WinStay}/10000; 
\text{WinSwitch} \leftarrow \text{WinSwitch}/10000$
Simulation Solution

Set WinStay and WinSwitch = 0.
Repeat 10,000 times
Draw $u \sim U(0, 1)$
If $u > 0.33$
Then WinSwitch ← WinSwitch + 1
Else WinStay ← WinStay + 1
WinStay ← WinStay/10000;
WinSwitch ← WinSwitch/10000

- Running this simulation, for instance, gives WinStay = .3328 and WinSwitch = .6672.
- The chance that the door you first picked is a winner is 1/3.
- In this case, just constructing the simulation gives the answer.
Simulation Solution

Set WinStay and WinSwitch = 0.
Repeat 10,000 times
Draw $u \sim U(0, 1)$
If $u > 0.33$
Then WinSwitch ← WinSwitch + 1
Else WinStay ← WinStay + 1
WinStay ← WinStay/10000;
WinSwitch ← WinSwitch/10000

- Running this simulation, for instance, gives WinStay = .3328 and WinSwitch = .6672.
- The chance that the door you first picked is a winner is 1/3.
- In this case, just constructing the simulation gives the answer.
- The problem with the previous formulation:
  $P(C^c|A) = 1/2$
More Reasons for Simulations

- Too many input variables to conduct a physical experiment (e.g., cancer treatment may involve a combination of treatment variables and exogenous variables that are too numerous for clinical trials).
More Reasons for Simulations

- Too many input variables to conduct a physical experiment (e.g., cancer treatment may involve a combination of treatment variables and exogenous variables that are too numerous for clinical trials).
- Ethical reasons may prevent physical experiments. For example, we cannot crash cars with actual people in them to test designs for safety.
More Reasons for Simulations

- Too many input variables to conduct a physical experiment (e.g., cancer treatment may involve a combination of treatment variables and exogenous variables that are too numerous for clinical trials).
- Ethical reasons may prevent physical experiments. For example, we cannot crash cars with actual people in them to test designs for safety.
- Economic considerations prevent physical simulations. For example, we cannot manufacture multiple wings for the space shuttle and hit them with foam at velocity to understand the sensitivity to foam strikes during launch.
Finite Element Simulations

- Finite element models (FEM) are a popular tool in determining the reliability of complex structures.
Finite Element Simulations

- Finite element models (FEM) are a popular tool in determining the reliability of complex structures.
- FEM are a low cost alternative to building prototypes and conducting physical experimentation.
Finite Element Simulations

- Finite element models (FEM) are a popular tool in determining the reliability of complex structures.
- FEM are a low cost alternative to building prototypes and conducting physical experimentation.
- FEM were used to study impacts from foam.
Finite Element Models of Impact

- Experiments and Simulations
- Models for Simulation Output
- Design of Simulation Experiments
- Approaches to Prediction from Simulations
- Example Learning from Simulations

800 ft/sec
Crack

850 ft/sec
Crack / Hole
Agent-Based Simulations

- Agent-based simulations have become increasingly important to understand complexity with a large numbers of interacting components.
Agent-Based Simulations

- Agent-based simulations have become increasingly important to understand complexity with a large numbers of interacting components.
- Agent-based simulations also enable emergent behavior.
Agent-Based Simulations

- Agent-based simulations have become increasingly important to understand complexity with a large numbers of interacting components.
- Agent-based simulations also enable emergent behavior.
- We have built and applied agent-based simulations in a number of areas to include planning for earthquakes, military operations, and oil spills.
Agent-Based Simulations

- Agent-based simulations have become increasingly important to understand complexity with a large numbers of interacting components.
- Agent-based simulations also enable emergent behavior.
- We have built and applied agent-based simulations in a number of areas to include planning for earthquakes, military operations, and oil spills.
- Disaster Response Information Flow and Technology Simulation (DRIFTS) combines agent-based modeling with a geographic information system (GIS) to yield a computer simulation with a high enough level of resolution to simulate movement of individuals in a spatially accurate environment.
Casualties from HAZUS with DRIFTS Laydown
DRIFTS Police Motion
Goals for Knowledge Discovery from Simulations

- Learn relationships between variables in complex problems. Possibly learn causality.
Goals for Knowledge Discovery from Simulations

- Learn relationships between variables in complex problems. Possibly learn causality.
- Optimize settings for input control variables.
## Goals for Knowledge Discovery from Simulations

- Learn relationships between variables in complex problems. Possibly learn causality.
- Optimize settings for input control variables.
- Adapt or evolve processes, products, and portfolios with changing conditions.
Simulation Variables

- Control variables, $x_c$, the variables that can be set by the engineer to affect the output.
Simulation Variables

- Control variables, $x_c$, the variables that can be set by the engineer to affect the output.
- Environmental or exogenous variables, $x_e$, variables that may affect the output but are not under the control of the engineers.
Simulation Variables

- Control variables, $x_c$, the variables that can be set by the engineer to affect the output.
- Environmental or exogenous variables, $x_e$, variables that may affect the output but are not under the control of the engineers.
- Model parameters, $x_m$, the physical constants or other characteristics of the model not known to the engineer for the simulated problem.
Simulation Variables

- Control variables, $x_c$, the variables that can be set by the engineer to affect the output.
- Environmental or exogenous variables, $x_e$, variables that may affect the output but are not under the control of the engineers.
- Model parameters, $x_m$, the physical constants or other characteristics of the model not known to the engineer for the simulated problem.
- We use $x = (x_c, x_e, x_m)$ to capture appropriate combinations of all three possible variables. Not all of them have to be used in a simulation.
Simulation Variables

- Control variables, $x_c$, the variables that can be set by the engineer to affect the output.
- Environmental or exogenous variables, $x_e$, variables that may affect the output but are not under the control of the engineers.
- Model parameters, $x_m$, the physical constants or other characteristics of the model not known to the engineer for the simulated problem.
- We use $x = (x_c, x_e, x_m)$ to capture appropriate combinations of all three possible variables. Not all of them have to be used in a simulation.
- The output of the simulation is represented by $y(x)$. This output may be vector valued.
Consider a real-valued output, $y(x)$, that is to be evaluated at training points, $x_1, x_2, \ldots, x_n$. 
Research Problem Formulation

- Consider a real-valued output, $y(x)$, that is to be evaluated at training points, $x_1, x_2, \ldots, x_n$.
- We want to learn a predictor of the output, $\hat{y}(x)$. 

Experiments and Simulations
Models for Simulation Output
Design of Simulation Experiments
Approaches to Prediction from Simulations
Example Learning from Simulations
Consider a real-valued output, $y(x)$, that is to be evaluated at training points, $x_1, x_2, \ldots, x_n$.

We want to learn a predictor of the output, $\hat{y}(x)$.

Judge performance with integrated squared error

$$\int_{\mathcal{X}} [y(x) - \hat{y}(x)]^2 w(x) dx$$
Sample Research Problems

- Optimization: find extreme values of $y(x)$ or find

$$\text{argmax}[y(x)]$$

for $x \in X$
Sample Research Problems

- Optimization: find extreme values of $y(x)$ or find
  \[ \text{argmax}[y(x)] \]
  for $x \in \mathcal{X}$

- Understanding: How is the variance in $x_e$ expressed in $y(x)$?
Sample Research Problems

- **Optimization**: find extreme values of $y(x)$ or find
  $$\text{argmax}[y(x)]$$
  for $x \in \mathcal{X}$

- **Understanding**: How is the variance in $x_e$ expressed in $y(x)$?

- **Calibration**: Find a fit between the simulation output and the known output from some typically physical process. This means finding the $x_m$ that enable this best fit in terms of integrated squared error.
Consider the random function $Y(x)$ with domain $\mathcal{X}$ and realization $y(x)$. 

Consider the random function $Y(x)$ with domain $\mathcal{X}$ and realization $y(x)$.

Can make the random process explicit with an outcome space $\Omega$. So $y(x) = Y(x, \omega)$ for $\omega \in \Omega$. 
Stochastic Process Model

- Consider the random function $Y(x)$ with domain $\mathcal{X}$ and realization $y(x)$.
- Can make the random process explicit with an outcome space $\Omega$. So $y(x) = Y(x, \omega)$ for $\omega \in \Omega$.
- For many problems we consider the environmental variables as stochastic with known distributions implemented in the simulations, so the stochastic process is $Y(x_c, X_e)$, where the $X_e$ have known distributions.
Best MSPE Predictor

- Suppose we want to predict the random response, $Y$, using the training data, $Y_1, \ldots, Y_n$ obtained from $n$ runs of the simulation.
Best MSPE Predictor

- Suppose we want to predict the random response, $Y$, using the training data, $Y_1, \ldots, Y_n$ obtained from $n$ runs of the simulation.
- A reasonable criterion to use is Mean Square Prediction Error (MPSE).

$$MSPE \equiv E[(Y - \hat{Y})^2]$$
Best MSPE Predictor

- Suppose we want to predict the random response, $Y$, using the training data, $Y_1, \ldots, Y_n$ obtained from $n$ runs of the simulation.
- A reasonable criterion to use is Mean Square Prediction Error (MPSE).

$$\text{MSPE} \equiv E[(Y - \hat{Y})^2]$$

- Theorem: Suppose $(Y, Y_1, \ldots, Y_n)$ has a joint distribution for which the conditional expectation of $Y$ given $Y_1, \ldots, Y_n$ exists. Then the best MSPE for $Y$ is

$$\hat{Y} = E(Y|Y_1, \ldots, Y_n)$$
DOE Considerations

- As noted traditional physical experiments are concerned with randomization, blocking, and replication.
DOE Considerations

- As noted traditional physical experiments are concerned with randomization, blocking, and replication.
- Also, multicollinearity is controlled by orthogonal or nearly orthogonal designs.
DOE Considerations

- As noted traditional physical experiments are concerned with randomization, blocking, and replication.
- Also, multicollinearity is controlled by orthogonal or nearly orthogonal designs.
- Designs should enable diagnostics to detect lack of fit, particularly lack of linearity.
DOE Considerations

- As noted traditional physical experiments are concerned with randomization, blocking, and replication.
- Also, multicollinearity is controlled by orthogonal or nearly orthogonal designs.
- Designs should enable diagnostics to detect lack of fit, particularly lack of linearity.
- For simulations, runs at exactly the same points yield the same answer. Uncertainty derives from a lack of knowledge of the functional relationship between inputs and response.
DOE Considerations

- As noted traditional physical experiments are concerned with randomization, blocking, and replication.
- Also, multicollinearity is controlled by orthogonal or nearly orthogonal designs.
- Designs should enable diagnostics to detect lack of fit, particularly lack of linearity.
- For simulations, runs at exactly the same points yield the same answer. Uncertainty derives from a lack of knowledge of the functional relationship between inputs and response.
- This implies we should conduct the experiments over as broad a range of the input space as possible.
Classical DOE

- Classical DOE builds on a factorial representation of the experimental inputs.
Classical DOE

- Classical DOE builds on a factorial representation of the experimental inputs.
- Factorial designs explore the vertices of the design space.
Classical DOE builds on a factorial representation of the experimental inputs.

Factorial designs explore the vertices of the design space.

Extensions, such as central composite and Box-Behnken enable broad explorations to include tests for nonlinearities.
In simulations our goal is to obtain a set of inputs for the process (simulation) that span or fill the design space.
In simulations our goal is to obtain a set of inputs for the process (simulation) that span or fill the design space.

Find good space filling experimental designs by minimizing discrepancy (supremum of the absolute difference of the proportion of points in subspaces to fractional hyper-volumes of those subspaces)

\[
L = \sup_{C^m} \left| \frac{N(b, P)}{n} - \omega(0, b) \right|
\]
Latin Hypercube Designs

- Consider a unit square for the design space, $[0, 1]^2$. 
Latin Hypercube Designs

- Consider a unit square for the design space, $[0, 1]^2$.
- For an $n$ point design divide the design space into $n$ equally spaced intervals.
Latin Hypercube Designs

- Consider a unit square for the design space, $[0, 1]^2$.
- For an $n$ point design divide the design space into $n$ equally spaced intervals.
- Fill these cells with integers, $1, \ldots, n$ to form a Latin square.
Latin Hypercube Designs

- Consider a unit square for the design space, $[0, 1]^2$.
- For an $n$ point design divide the design space into $n$ equally spaced intervals.
- Fill these cells with integers, 1, . . . , $n$ to form a Latin square.
- Instead of a Latin square we can use a Halton sequence (or other).
Latin Hypercube Designs

- Consider a unit square for the design space, \([0, 1]^2\).
- For an \(n\) point design divide the design space into \(n\) equally spaced intervals.
- Fill these cells with integers, \(1, \ldots, n\) to form a Latin square.
- Instead of a Latin square we can use a Halton sequence (or other).
- We tested Latin Hypercube Sampling (LHS), Halton Sequence Sampling (HSS), and a combination of both.
Example LHS and HSS

Experiments and Simulations
Models for Simulation Output
Design of Simulation Experiments
Approaches to Prediction from Simulations
Example Learning from Simulations
Sampling Results

Experiments and Simulations
- Models for Simulation Output
- Design of Simulation Experiments
- Approaches to Prediction from Simulations
- Example Learning from Simulations

HSS v. LHS Discrepancy (p=3)

HSS v. LHS Discrepancy (p=4)

HSS v. LHS Discrepancy (p=5)

HSS v. LHS Discrepancy (p=6)
Methods from Knowledge Discovery and Data Mining

- Linear models
- Generalized additive models
- CART
- Neural networks
- Polynomial regression
- Radial basis functions
- Kriging

- LOESS
- Projection pursuit regression
- Instance based regression
- Support vector machines
- Logistic regression
- Principal component regression
Just Apply Knowledge Discovery and Data Mining?

- Without a formal process model, applying data mining is equivalent to data dredging! You will find patterns but are the patterns real?
Just Apply Knowledge Discovery and Data Mining?

- Without a formal process model, applying data mining is equivalent to data dredging! You will find patterns but are the patterns real?
- Simulations can generate arbitrary levels of significance. You can fool yourself and others!
Just Apply Knowledge Discovery and Data Mining?

- Without a formal process model, applying data mining is equivalent to data dredging! You will find patterns but are the patterns real?
- Simulations can generate arbitrary levels of significance. You can fool yourself and others!
- To learn from simulations, the simulations must be calibrated.
Just Apply Knowledge Discovery and Data Mining?

- Without a formal process model, applying data mining is equivalent to data dredging! You will find patterns but are the patterns real?
- Simulations can generate arbitrary levels of significance. You can fool yourself and others!
- To learn from simulations, the simulations must be calibrated.
- The predictive approach should be in the context of response surface methodology (RSM) to really learn from simulations.
Response Surface Methodology

Box and Draper, 1987
Example Response Surfaces
Advantages of RSM

- Employs the DOE to enable learning from simulations.
Advantages of RSM

- Employs the DOE to enable learning from simulations.
- Enables specification of all variables and interactions.
Advantages of RSM

- Employs the DOE to enable learning from simulations.
- Enables specification of all variables and interactions.
- In iterative form RSM allows for evolution or optimization of the process.
A Modified RSM for Simulations

Experiments and Simulations
Models for Simulation Output
Design of Simulation Experiments
Approaches to Prediction from Simulations
Example Learning from Simulations

Predictive Technology Lab
Advantages for MRSM for Simulations

- Better for problems with less certainty and less a priori information.
Advantages for MRSM for Simulations

- Better for problems with less certainty and less a priori information.
- Handles a larger decision space with more variables.
Advantages for MRSM for Simulations

- Better for problems with less certainty and less a priori information.
- Handles a larger decision space with more variables.
- Handles more complex relationships.
Advantages for MRSM for Simulations

- Better for problems with less certainty and less a priori information.
- Handles a larger decision space with more variables.
- Handles more complex relationships.
- More likely to find areas of true optimality.
Advantages for MRSM for Simulations

- Better for problems with less certainty and less a priori information.
- Handles a larger decision space with more variables.
- Handles more complex relationships.
- More likely to find areas of true optimality.
- Uses complementary data mining techniques to better capture relationships and to guard against error.
Additive Models for Prediction

- Additive models, such as Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS) distinguish well between contributing and noise variables.
Additive Models for Prediction

- Additive models, such as Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS) distinguish well between contributing and noise variables.
- Can perform feature selection and understanding.
Additive Models for Prediction

- Additive models, such as Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS) distinguish well between contributing and noise variables.
- Can perform feature selection and understanding.
- Computation times are fast for high-dimensional problems.
Additive Models for Prediction

- Additive models, such as Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS) distinguish well between contributing and noise variables.
- Can perform feature selection and understanding.
- Computation times are fast for high-dimensional problems.
- Perform well on problems where the curse of dimensionality would likely create problems for other techniques.
Additive Models for Prediction

- Additive models, such as Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS) distinguish well between contributing and noise variables.
- Can perform feature selection and understanding.
- Computation times are fast for high-dimensional problems.
- Perform well on problems where the curse of dimensionality would likely create problems for other techniques.
- Variables do not contribute globally across the response surface.
Additive Models for Prediction

- Additive models, such as Generalized Additive Models (GAM) and Multivariate Adaptive Regression Splines (MARS) distinguish well between contributing and noise variables.
- Can perform feature selection and understanding.
- Computation times are fast for high-dimensional problems.
- Perform well on problems where the curse of dimensionality would likely create problems for other techniques.
- Variables do not contribute globally across the response surface.
- MARS uses simple linear splines in reflected pairs.
Reflected Pairs

\[
(X - t)_+ = \begin{cases} 
  x - t & \text{if } x > t \\
  0 & \text{else}
\end{cases}
\]

\[
(t - X)_+ = \begin{cases} 
  t - x & \text{if } t > x \\
  0 & \text{else}
\end{cases}
\]
These spline functions are used in additive models:

\[ f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \]
MARS Additive Models

These spline functions are used in additive models:

\[ f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \]

The models are built up with the reflected pairs that best reduce the error. Model selection is accomplished using generalized cross validation.

\[ GCV = \frac{\sum_{i=1}^{n} (y_i - \hat{f}_\lambda(x_i))^2}{(1 - \frac{M(\lambda)}{n})^2} \]
MARS Additive Models

These spline functions are used in additive models:

\[ f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \]

The models are built up with the reflected pairs that best reduce the error. Model selection is accomplished using generalized cross validation.

\[ GCV = \frac{\sum_{i=1}^{n} (y_i - \hat{f}_\lambda(x_i))^2}{(1 - \frac{M(\lambda)}{n})^2} \]

In this formula, \( \lambda \) is the number of terms in the model. \( M(\lambda) \) is the effective number of parameters in the model since the knots used in the splines are counted as additional parameters.
MARS Approximation

Experiments and Simulations
Models for Simulation Output
Design of Simulation Experiments
Approaches to Prediction from Simulations
Example Learning from Simulations
True surface

MARS approximation
Pan and Zoom Search

Pan

Zoom

Pan & Zoom

$X^{(0)} - X^{(1)}$
Test Cases

Experiments and Simulations
Models for Simulation Output
Design of Simulation Experiments
Approaches to Prediction from Simulations
Example Learning from Simulations

Six-Hump Camel Back Function
(SC, -1.032, 6)

Basin Function
(BA, 0, 15)

Branin Function
(BR, 0.398, 3)

Generalized Polynomial Function
(GF, 0, 5)

Peaks Function
(PE, -6.551, 3)

Rastrigin Function
(RS, -2, 50)
Test Preparation

- Clean data: Only significant variables are included in the data set.
Test Preparation

- **Clean data**: Only significant variables are included in the data set.
- **Noisy data**: Eight insignificant random variables are added creating a low signal to noise ratio data set (2:8, 6:8).
Test Preparation

- **Clean data:** Only significant variables are included in the data set.
- **Noisy data:** Eight insignificant random variables are added creating a low signal to noise ratio data set (2:8, 6:8).
- **Results are an average of:** 100 iterations when LHS (stochastic) is used single iteration when HSS (deterministic) is used.
<table>
<thead>
<tr>
<th>Comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Global Optimization (MARS/RSM v. GA, SA)</strong></td>
</tr>
</tbody>
</table>

- Experiments and Simulations
- Models for Simulation Output
- Design of Simulation Experiments
- Approaches to Prediction from Simulations
- Example Learning from Simulations
## Comparisons

- Global Optimization (MARS/RSM v. GA, SA)
- Function approximation (MARS v. GAM, NN)
Comparisons

- Global Optimization (MARS/RSM v. GA, SA)
- Function approximation (MARS v. GAM, NN)
- Variable screening
Global Optimization Results

<table>
<thead>
<tr>
<th>Test Function</th>
<th>Genetic Algorithms</th>
<th>MARS/RSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eval</td>
<td>Rel. Error</td>
</tr>
<tr>
<td>SC</td>
<td>1683</td>
<td>0.48</td>
</tr>
<tr>
<td>BR</td>
<td>2132</td>
<td>1.56</td>
</tr>
<tr>
<td>GF</td>
<td>2288</td>
<td>4.85</td>
</tr>
<tr>
<td>RS</td>
<td>1845</td>
<td>0.43</td>
</tr>
<tr>
<td>BA</td>
<td>1285</td>
<td>0.50</td>
</tr>
<tr>
<td>PE</td>
<td>1245</td>
<td>0.52</td>
</tr>
<tr>
<td>HE</td>
<td>2500*</td>
<td>20.11</td>
</tr>
</tbody>
</table>
### Simulated Annealing Results

<table>
<thead>
<tr>
<th>Test Function</th>
<th>Simulated Annealing</th>
<th>MARS/RSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eval</td>
<td>Rel. Error</td>
</tr>
<tr>
<td>SC</td>
<td>3361</td>
<td>0.59</td>
</tr>
<tr>
<td>BR</td>
<td>1074</td>
<td>0.46</td>
</tr>
<tr>
<td>GF</td>
<td>541</td>
<td>22.87</td>
</tr>
<tr>
<td>RS</td>
<td>708</td>
<td>9.57</td>
</tr>
<tr>
<td>BA</td>
<td>487</td>
<td>4.59</td>
</tr>
<tr>
<td>PE</td>
<td>1042</td>
<td>0.57</td>
</tr>
<tr>
<td>HE</td>
<td>6565</td>
<td>65.60</td>
</tr>
</tbody>
</table>
### Function Approximation Results

<table>
<thead>
<tr>
<th>Function</th>
<th>Sample Size</th>
<th>MARS</th>
<th>GAM</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>50</td>
<td>9.69E+04</td>
<td>3.38E+06</td>
<td>3.64E+07</td>
</tr>
<tr>
<td>BR</td>
<td>50</td>
<td>4.61E+01</td>
<td>5.59E+03</td>
<td>7.98E+03</td>
</tr>
<tr>
<td>GF</td>
<td>50</td>
<td>2.56E+08</td>
<td>2.91E+09</td>
<td>3.15E+10</td>
</tr>
<tr>
<td>RS</td>
<td>100</td>
<td>8.69</td>
<td>10.02</td>
<td>28.28</td>
</tr>
<tr>
<td>BA</td>
<td>50</td>
<td>0.25</td>
<td>0.77</td>
<td>29.26</td>
</tr>
<tr>
<td>PE</td>
<td>50</td>
<td>23.96</td>
<td>45.64</td>
<td>242.34</td>
</tr>
<tr>
<td>HE</td>
<td>50</td>
<td>1.15E+03</td>
<td><strong>1.51E+02</strong></td>
<td>1.21E+03</td>
</tr>
</tbody>
</table>

**Clean data**

<table>
<thead>
<tr>
<th>Function</th>
<th>Sample Size</th>
<th>MARS</th>
<th>GAM</th>
<th>NN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>50</td>
<td>1.07E+06</td>
<td><strong>2.21E+04</strong></td>
<td>3.04E+07</td>
</tr>
<tr>
<td>BR</td>
<td>50</td>
<td><strong>5.70E+03</strong></td>
<td>8.03E+04</td>
<td>6.77E+04</td>
</tr>
<tr>
<td>GF</td>
<td>50</td>
<td>7.96E+10</td>
<td><strong>1.40E+10</strong></td>
<td>3.17E+10</td>
</tr>
<tr>
<td>RS</td>
<td>100</td>
<td><strong>25.50</strong></td>
<td>63.09</td>
<td>45.01</td>
</tr>
<tr>
<td>BA</td>
<td>50</td>
<td>1.93</td>
<td>6.00</td>
<td>36.49</td>
</tr>
<tr>
<td>PE</td>
<td>50</td>
<td>64.59</td>
<td>84.09</td>
<td>83.16</td>
</tr>
<tr>
<td>HE</td>
<td>50</td>
<td><strong>1.05E+02</strong></td>
<td>1.60E+04</td>
<td>7.92E+04</td>
</tr>
</tbody>
</table>

**Noisy data**
Variable Screening Results

<table>
<thead>
<tr>
<th>Function</th>
<th>Sample Size</th>
<th>Predictors</th>
<th>Noise Variables</th>
<th>HSS Significant Variables</th>
<th>LHS Significant Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC</td>
<td>50</td>
<td>$X_1, X_2$</td>
<td>$X_3 - X_{10}$</td>
<td>$X_1, X_2, X_4, X_5, X_6$</td>
<td>$X_1, X_2, X_6, X_9, X_{10}$</td>
</tr>
<tr>
<td>BR</td>
<td>50</td>
<td>$X_1, X_2$</td>
<td>$X_3 - X_{10}$</td>
<td>$X_1, X_2$</td>
<td>$X_1, X_2, X_{10}$</td>
</tr>
<tr>
<td>GF</td>
<td>50</td>
<td>$X_1, X_2$</td>
<td>$X_3 - X_{10}$</td>
<td>$X_1, X_2, X_3, X_6, X_7, X_{10}$</td>
<td>$X_1, X_2, X_5, X_6$</td>
</tr>
<tr>
<td>RS</td>
<td>100</td>
<td>$X_1, X_2$</td>
<td>$X_3 - X_{10}$</td>
<td>$X_1, X_2, X_4, X_5, X_7$</td>
<td>$X_2, X_5, X_9$</td>
</tr>
<tr>
<td>BA</td>
<td>50</td>
<td>$X_1, X_2$</td>
<td>$X_3 - X_{10}$</td>
<td>$X_1, X_2, X_4, X_5$</td>
<td>$X_1, X_2$</td>
</tr>
<tr>
<td>PE</td>
<td>50</td>
<td>$X_1, X_2$</td>
<td>$X_3 - X_{10}$</td>
<td>$X_1, X_2, X_5, X_6$</td>
<td>$X_1, X_2$</td>
</tr>
<tr>
<td>HE</td>
<td>50</td>
<td>$X_1, X_6$</td>
<td>$X_7 - X_{14}$</td>
<td>$X_3 - X_6$</td>
<td>$X_1, X_6, X_7, X_9, X_{14}$</td>
</tr>
</tbody>
</table>

*MARS found all but one of the proper predictors*
High Level Steps in Examples

- Use a space-filling design of experiment (DOE) to obtain sample observations, $x$, across the domain space, $\mathcal{X}$. 
High Level Steps in Examples

- Use a space-filling design of experiment (DOE) to obtain sample observations, \( x \), across the domain space, \( \mathcal{X} \).
- Create an approximation (model) using MARS to estimate the response function, \( \hat{y}(x) \).
High Level Steps in Examples

- Use a space-filling design of experiment (DOE) to obtain sample observations, $x$, across the domain space, $\mathcal{X}$.
- Create an approximation (model) using MARS to estimate the response function, $\hat{y}(x)$.
- Use the model to predict good region(s) for follow-on simulation.
High Level Steps in Examples

- Use a space-filling design of experiment (DOE) to obtain sample observations, $x$, across the domain space, $\mathcal{X}$.
- Create an approximation (model) using MARS to estimate the response function, $\hat{y}(x)$.
- Use the model to predict good region(s) for follow-on simulation.
- Iterate until change in the response is sufficiently small.
High Level Steps in Examples

- Use a space-filling design of experiment (DOE) to obtain sample observations, \( x \), across the domain space, \( \mathcal{X} \).
- Create an approximation (model) using MARS to estimate the response function, \( \hat{y}(x) \).
- Use the model to predict good region(s) for follow-on simulation.
- Iterate until change in the response is sufficiently small.
- See [2] for more details on these examples. In addition to these examples we have also developed simulations to evaluate and optimize airborne surveillance planning [3], and oil spill response planning [1].
Vehicle Frontal Crash

Minimize Head Injury Criterion (HIC)

\[
HIC = \max \left\{ \left[ \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} a(t) dt \right]^2 \cdot 5(t_2 - t_1) \right\}
\]

Subject to

- Intrusion \( \leq 550mm \)
- \( t_{hood} \in [1, 5]mm \)
- \( t_{bumper} \in [1, 5]mm \)
Vehicle Crash Simulation in LS Dyna

Experiments and Simulations
Models for Simulation Output
Design of Simulation Experiments
Approaches to Prediction from Simulations
Example Learning from Simulations
Vehicle Crash Response Variables

- HIC, Intrusion 1 and Intrusion 2.
## Vehicle Crash Response Variables

- HIC, Intrusion 1 and Intrusion 2.
- LS-OPT in LS-Dyna had three optimization techniques for use on this problem: Polynomial regression, Kriging, and Neural Networks.
Vehicle Crash Response Variables

- HIC, Intrusion 1 and Intrusion 2.
- LS-OPT in LS-Dyna had three optimization techniques for use on this problem: Polynomial regression, Kriging, and Neural Networks.
- MRSM as a learning method from simulations was compared with all LS-OPT techniques.
Vehicle Frontal Crash Results

Comparison of Approximation Procedures for Minimizing HIC

- MARS
- NN
- Kriging
- PR

HIC vs. Function evaluations
# Constraint Violation

<table>
<thead>
<tr>
<th></th>
<th>Eval</th>
<th>HIC</th>
<th>Intru</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARS/RSM</td>
<td>50</td>
<td>128.7</td>
<td>547.92</td>
</tr>
<tr>
<td>PR</td>
<td>51</td>
<td>128.9</td>
<td>549.8</td>
</tr>
<tr>
<td>NN</td>
<td>51</td>
<td>126.8</td>
<td>550.54</td>
</tr>
<tr>
<td>Kriging</td>
<td>81</td>
<td>129.7</td>
<td>550.41</td>
</tr>
</tbody>
</table>
6 Dimensional Vehicle Crash Problem

- The finite element model simulates a vehicle crashing into a stationary pole resulting in 5 measured responses.
6 Dimensional Vehicle Crash Problem

- The finite element model simulates a vehicle crashing into a stationary pole resulting in 5 measured responses.
- The objective is to minimize the total mass of the parts subject to intrusion, HIC and thickness constraints.
Finite Element Simulation for Frontal Crash

Experiments and Simulations

Models for Simulation Output

Design of Simulation Experiments

Approaches to Prediction from Simulations

Example Learning from Simulations
Vehicle Crash Problem Formulation

Minimize Mass
Subject to

\[
\text{Intrusion} < 550mm \\
(HIC) < 900mm \\
t_i \in [1, 6]mm \\
i \in 1, \ldots, 6
\]

Where the \( t_i \) represent bumper, grill, hood, rail back, rail front, and roof.
### 6 Dimension Crash Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Eval</th>
<th>HIC</th>
<th>Intru</th>
<th>Mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARS/ RSM</td>
<td>59</td>
<td>906.2</td>
<td>504.9</td>
<td>2.6147</td>
</tr>
<tr>
<td>QR</td>
<td>223</td>
<td>875.7</td>
<td>512.7</td>
<td>2.6305</td>
</tr>
<tr>
<td>LR*</td>
<td>52</td>
<td>888.9</td>
<td>499.6</td>
<td>2.740</td>
</tr>
<tr>
<td>NN</td>
<td>74</td>
<td>908.4</td>
<td>513.0</td>
<td>2.5956</td>
</tr>
<tr>
<td>Kriging</td>
<td>50</td>
<td>889.6</td>
<td>511.5</td>
<td>2.7992</td>
</tr>
</tbody>
</table>
6 Dimension Design Conclusions

- Only one of the over 500 points sampled met all the constraints simultaneously (found using LR).
6 Dimension Design Conclusions

- Only one of the over 500 points sampled met all the constraints simultaneously (found using LR).
- If a designer wishes to maximize the number of design alternatives that both minimize the Mass and minimize violation of the Intrusion and HIC constraints in a timely manner while offering an interpretable and accurate response surface approximation, then MRSM with MARS (or a similar additive model) is the best choice.
Automobile Hood Design

- 30% of pedestrian deaths caused by head-hood impacts.
Automobile Hood Design

- 30% of pedestrian deaths caused by head-hood impacts.
- The most relevant characteristics for a safer hood design are: sufficient stiffness to allow a low under hood clearance; and, an energy absorbing structure that effectively decelerates the head impact (Untaroiu et al., 2006).
Automobile Hood Design

- 30% of pedestrian deaths caused by head-hood impacts.
- The most relevant characteristics for a safer hood design are: sufficient stiffness to allow a low under hood clearance; and, an energy absorbing structure that effectively decelerates the head impact (Untaroiu et al., 2006).
- These two objectives are conflicting and must be resolved through experiments. Additional constraints on the spool and panel design variables add to the complexity of the problem.
30% of pedestrian deaths caused by head-hood impacts.

The most relevant characteristics for a safer hood design are: sufficient stiffness to allow a low under hood clearance; and, an energy absorbing structure that effectively decelerates the head impact (Untaroiu et al., 2006).

These two objectives are conflicting and must be resolved through experiments. Additional constraints on the spool and panel design variables add to the complexity of the problem.

The impact of the head on the hood must be measured and constrained.
Hood Design Through Simulation

- Finite element models provide the most effective approach to discovering good hood finite element analytic models currently 10-20 hrs/run.
Hood Design Through Simulation

- Finite element models provide the most effective approach to discovering good hood finite element analytic models currently 10-20 hrs/run.
- HIC cannot be derived analytically
Hood Design Through Simulation

- Finite element models provide the most effective approach to discovering good hood finite element analytic models currently 10-20 hrs/run.
- HIC cannot be derived analytically
- The design objective is to minimize total stroke which is the highest normal deformation of the lower panel (S) plus the total thickness of the hood, H.
Use an LS-DYNA analytic model in accordance with ISO and IHRA specifications.
Scull Model in LS Dyna

FE Development

validation of skin

Based on JARI data (Matsui et al 2003, 2005) and in agreement with requirements of ISO, IHRA/Japan MLIT proposals

To solve this problem requires another LS Dyna model of the scull.
Complete Hood Design

Minimize \( \text{Total\_stroke} \)

Subject to:

\[ \begin{align*}
HIC & < 1000 \\
H & < 40 \, \text{mm} \\
4 & \leq \text{radius} \leq 10 \\
5 & \leq h \leq 30 \\
5 & \leq \text{rad\_int} \leq 20 \\
0.08 & \leq \text{th\_cyl} \leq 0.2 \\
0.6 & \leq \text{th\_up} \leq 2 \\
0.6 & \leq \text{th\_down} \leq 2
\end{align*} \]
Hood MRSM

Experiments and Simulations

Models for Simulation Output

Design of Simulation Experiments

Approaches to Prediction from Simulations

Example Learning from Simulations

HIC fit \( R^2 = .974, .989 \)

Total_stroke fit \( R^2 = .986, .978 \)

S. Crino

Dissertation Defense
Hood MARS Fit

Experiments and Simulations

Models for Simulation Output

Design of Simulation Experiments

Approaches to Prediction from Simulations

Example Learning from Simulations

HIC fit $R^2_a = .974, .989$

Total_stroke fit $R^2_a = .986, .978$

Contradiction

S. Crino

Dissertation Defense
Hood Design Results

<table>
<thead>
<tr>
<th>Parameters</th>
<th>radius</th>
<th>h</th>
<th>rad_int</th>
<th>th_cyl</th>
<th>th_up</th>
<th>th_down</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
<td><strong>Min</strong></td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>0.08</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td><strong>Max</strong></td>
<td>10</td>
<td>30</td>
<td>20</td>
<td>0.2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Best</strong></td>
<td>6</td>
<td>10</td>
<td>6</td>
<td>0.1</td>
<td>0.74</td>
<td>0.7</td>
</tr>
</tbody>
</table>

The spool hood design meets ISO and IHRA specifications based on MARS/RSM final solution.
Conclusions

- Cost, ethical considerations, and problem complexity motivate the use of simulations to understand cause and effect relationships and to optimize designs.
Conclusions

- Cost, ethical considerations, and problem complexity motivate the use of simulations to understand cause and effect relationships and to optimize designs.
- Although simulations are frequently less expensive than physical experiments; nonetheless they can be complex and require long run times.
Conclusions

- Cost, ethical considerations, and problem complexity motivate the use of simulations to understand cause and effect relationships and to optimize designs.
- Although simulations are frequently less expensive than physical experiments; nonetheless they can be complex and require long run times.
- To learn from simulations requires methods that effectively predict future simulation outputs with small numbers of runs.
Conclusions

- Cost, ethical considerations, and problem complexity motivate the use of simulations to understand cause and effect relationships and to optimize designs.
- Although simulations are frequently less expensive than physical experiments; nonetheless they can be complex and require long run times.
- To learn from simulations requires methods that effectively predict future simulation outputs with small numbers of runs.
- The Modified Response Surface Methodology combines space filling experimental designs and flexible nonlinear modeling with additive models, such as MARS.
Conclusions

- Cost, ethical considerations, and problem complexity motivate the use of simulations to understand cause and effect relationships and to optimize designs.
- Although simulations are frequently less expensive than physical experiments; nonetheless they can be complex and require long run times.
- To learn from simulations requires methods that effectively predict future simulation outputs with small numbers of runs.
- The Modified Response Surface Methodology combines space filling experimental designs and flexible nonlinear modeling with additive models, such as MARS.
- MRSM provides an approach to learning from simulations that leads to good solutions with modest numbers of simulation runs.
Conclusions

- Cost, ethical considerations, and problem complexity motivate the use of simulations to understand cause and effect relationships and to optimize designs.
- Although simulations are frequently less expensive than physical experiments; nonetheless they can be complex and require long run times.
- To learn from simulations requires methods that effectively predict future simulation outputs with small numbers of runs.
- The Modified Response Surface Methodology combines space filling experimental designs and flexible nonlinear modeling with additive models, such as MARS.
- MRSM provides an approach to learning from simulations that leads to good solutions with modest numbers of simulation runs.
- Testing with Finite Element simulations, as well as standard test sets shows the potential for Modified Response Surface Methodology to enable knowledge discovery from simulations.
References

