Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

MONOTONIC RESPONSE SURFACE ESTIMATION BY CONSTRAINED COEFFICIENTS

Frederick A. Ahrens

Operations Research
Raytheon Missile Systems
1151 E. Hermans Road, Bldg 805, M/S G4
Tucson, AZ 85756-9367, USA

ABSTRACT

Classic first and second-order response surface models (RSM) do not automatically observe monotonicity, while in many real problems, the researcher knows the response to be monotonic in some variables. This paper provides the constraints on coefficients that ensure monotonicity and offers some approaches for estimating monotonically constrained response surfaces.

1 INTRODUCTION

Multiple regression using linear models of the simulation response surface is a powerful way to reconstruct variable effects and interactions from limited observations. While many advanced data analysis methods are available, multiple regression using first or second-order response surfaces are widely taught and recommended (Penn State 2014; Xu, H. 2013; SAS Institute 2014; ReliaSoft Corporation 2013; NIST/SEMATECH April 2012).

A linear model of a response y over a multi-dimensional independent variable x has the form

$$y = f(x) + \varepsilon;\ f(x) = \sum_{i=1}^{m} a_i \phi_i(x), \tag{1}$$

where $\phi_i(x)$ are pre-determined scalar functions of x, ε is sampling error from the simulation, and the a_i are the unknown regression coefficients. The form of $f(x)$ is not guaranteed monotonic in any dimension of x, even if the functions $\phi_i(x)$ are monotonic, because the coefficients are allowed to be positive or negative.

To illustrate, consider a function of five dichotomous variables

$$f(A,B,C,D,E) = 1.650 + (0.875)A + (0.300)B + (0.175)C + (0.100)D + (0.050)E - (0.075)AB + (0.050)AC \tag{2}$$

where A, B, C, D and E all equal -1 or 1. Note that f is non-decreasing in all variables.

Let the function f represent ground truth, and suppose the researchers undertake to estimate it, using a 2^{5-1} fractional factorial experiment. However, researchers have prior knowledge that f is non-decreasing in the first three variables, A, B and C, while the directions of effects E and D remain unknown.

Table 1 shows the design of experiments (DOE) and simulation responses. The responses, which are single replicates, are contaminated by a sampling error which is normal with mean zero and standard deviation 0.3. Because the design in Table 1 has resolution V, it is possible to separate main effects and two-way interactions. Researchers choose to estimate the main effects plus interactions between the first four factors, namely, AB, AC, AD, BC, BD, CD, assuming that the interactions with E are negligible. A regression results in the estimated response

$$\hat{f}(A,B,C,D,E) = 1.551 + (0.954)A + (0.326)B + (0.108)C - (0.028)D + (0.077)E - (0.190)AB + (0.006)AC + (0.081)AD - (0.116)BC + (0.017)BD + (0.111)CD.$$

Table 1: 2^{5-1} fractional factorial design of experiments.

Experiment	A	B	C	D	E	Response
1	-1	-1	-1	-1	1	0.200
2	-1	-1	-1	1	-1	-0.255
3	-1	-1	1	-1	-1	0.243
4	-1	-1	1	1	1	0.137
5	-1	1	-1	-1	-1	1.127
6	-1	1	-1	1	1	0.909
7	-1	1	1	-1	1	1.255
8	-1	1	1	1	-1	1.162
9	1	-1	-1	-1	-1	2.016
10	1	-1	-1	1	1	2.045
11	1	-1	1	-1	1	2.622
12	1	-1	1	1	-1	2.797
13	1	1	-1	-1	1	2.989
14	1	1	-1	1	-1	2.517
15	1	1	1	-1	-1	2.183
16	1	1	1	1	1	2.873

The regression has a root mean square error of 0.264 (4 degrees of freedom). Figure 1 shows outputs of the prediction model for $A = E = -1$ side by side with the true response from (2). The estimate is non-monotonic in factor C, decreasing in C when $B = 1$ and $D = -1$, and increasing in other cases.

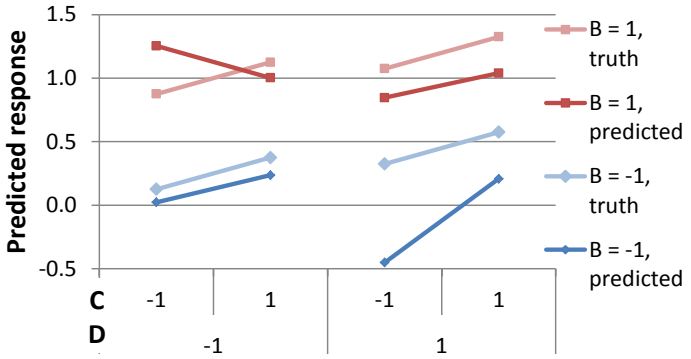

Figure 1: Predicted response from regression model (2) of response data (Table 1).

In this case, the corrupting sampling error together with a model space that admits non-monotonic functions, results in a non-monotonic prediction function even though the truth response is monotonic. The counter-intuitive effect is not a problem for statisticians who know that it is not statistically significant. However, Figure 1 can have problems in presentation to clients or across disciplines. Counter-intuitive results are a negative against credibility and can distract from the key points. Moreover, why not use a model space that is a more compliant representation of the studied processes if it were feasible? Prior research has found cases where monotonic models had better predictive power of monotonic processes than their unconstrained counterparts (Neelon and Dunson 2004).

In this paper, we will show how the functional form in (1) can be made monotonic in certain dimensions of x by applying a set of linear constraints on a_i, provided that the derivatives of $\phi_i(x)$ are bounded over the region of interest. One can estimate the coefficients by constrained least squares or

constrained maximum likelihood. Further, these models have a natural Bayesian approach, the constraints representing prior knowledge about the coefficients. Given prior probability densities on a_i whose support is restricted by linear constrains on a_i, and given the likelihood function for the sampling errors, we can generate random samples from the posterior distribution of a_i using one of the many Bayesian sampling algorithms.

Monotonic response surface models are of frequent interest in the literature. There appears to be little if any prior work on monotonic linear regression models, the direction of research being in non-parametric and semi-parametric models. The term "monotonic regression" traditionally refers to the pool adjacent violators (PAV) method in Ayer, et al. (1955). Rather than coefficients, the PAV directly estimates means of groups of observations that have ranks imposed on them. Originally, PAV applied only to functions of a single variable, but has since been extended to multiple variables using partial orderings on the observations (Barlow et al. 1972, Burdakow et al. 2004, Lim and Glynn 2006). In multivariate monotonic regression, the family of models consists of piecewise constant functions, which are of limited use in estimating between the observations.

Kay and Ungar (2000) give constraints on the coefficients of neural nets that ensure monotonicity of the neural net function and then solve a constrained least squares minimization. All of the examples are univariate, although the theory supports multivariate response surfaces. This paper applies essentially the same approach to second-order multivariate response surface models.

Hall and Huang (2001) found that kernel estimators can be made monotonic for one-dimensional functions. Racine and Parmeter (2008) extended the approach to multiple dimensions. In both the univariate and the multivariate cases, the existence of their solution depends on weak assumptions on the kernel functions over a bounded interval.

Bayesian approaches to monotonic response surface estimation are abundant. Gelfand and Kao (1991) use a Dirichlet mixture model over a family of monotonic functions to obtain Bayesian estimates of dosage response functions. If all of the functions in a family indexed by a parameter are monotonic, then so will be any mixture of functions from the family. Sampling the parameter against a Dirichlet process provides a prior distribution over a large class of monotonic functions. The concept was later applied to link functions for generalized linear models (GLMs), (Mallick and Gellfand 1994, Gelfand 1997). While these solutions only apply to responses of single variables, a Dirichlet mixture GLM could be combined with constrained coefficients, resulting in multivariate monotonic response surfaces. However, we restrict the present work to GLMs with fixed link functions and strictly linear models.

Researchers have restricted various response surface classes to subclasses of functions which are monotone in one or more dimensions. Neelon and Dunson (2004) propose a Bayesian monotonic regression using priors on the parameters of a piecewise linear spline function, the knot locations being among the parameters. Their extension to responses of several variables using a simple additive model of univariate functions ignores interactions between variables.

It is possible to constrain the derivatives of Gaussian processes (GP) at a finite set of points, thereby obtaining estimated response surfaces with acceptable behavior for the application. Riihimäki and Vehtari (2010) derive the constraints on GP weights for a user-selected set of points or "hints". The method does not guarantee global monotonicity, so two or more iterations with intermediate inspections may be necessary to reach an acceptable representation. Riihimäki and Vehtari demonstrate this method on a data set of 1,222 observations with seven independent variables, one of which is monotonic.

Lin and Dunson (2014) model a class of monotonic response surfaces as projections of GP, thereby obtaining a prior distribution over the class of monotonic functions. After drawing a Monte Carlo Markov Chain (MCMC) sample against the posterior of an ordinary GP, the estimator projects the sample paths onto the space of monotonic functions using a multivariate extension of PAV. They demonstrate the method on data sets of up to 1,024 observations with up to two independent variables.

This paper takes a simpler parametric approach than those just reviewed. Prior research has demonstrated the practical capability to match arbitrary monotonic shapes in one or two dimensions with

hundreds or thousands of design points. Often in simulation experimentation, designs of experiments (DOE) are relatively lean, the objective being to characterize a response against many variables for purposes of prediction or to assess uncertainty in the presence of incomplete knowledge of input variables. DOEs with as few as 10 unique cases per variable are common (Leoppky, Sacks and Welsh 2009). A parametric response surface may be needed to fill in the "white space" between observations. This paper will demonstrate estimation on a problem with 12 factors, 11 of which are known to be monotonic using the linear second-order response surface model

$$f(x) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} b_{ij}\, x_i x_j + \sum_{i=1}^{n} c_i x_i^2. \quad (3)$$

We derive the constraints for monotonicity for second-order response surface models. We discuss how they would be used in maximum likelihood and Bayesian estimation and offer some numerical examples. Section 2 will derive the monotonicity constraints first for model (1) and then for the more particular model (3). Section 3 will discuss statistical models based on (1) and (3). Section 4 will outline the Bayesian approach to estimation. Section 5 will present results of monotonically unconstrained and constrained models for data sets. Section 6 will summarize and provide some discussions of options and issues not addressed in the paper.

2 MONOTONIC CONSTRAINTS ON LINEAR MODELS

This section derives sufficient conditions on the coefficients a_i in (1), and on a_i, b_{ij} and c_i in (3) that ensure global monotonicity in one or more dimensions.

Suppose that $x = (x_1, x_2, \dots, x_n)$, $x \in A \subset \mathbb{R}^n$, and that the functions $\phi_i(x)$ for $i = 1, \dots, m$ are differentiable over A with bounded first partial derivatives. If monotonicity is required in several dimensions, then the constraints will be the union of the constraints over the monotonic dimensions. For now, assume that $f(x)$ shall be non-decreasing in x_1 for all $x \in A$. Choose L_{i1}, U_{i1} such that

$$L_{i1} \le \frac{\partial \phi_i}{\partial x_1} \le U_{i1}, \text{ for all } x \in A. \quad (4)$$

Define $M^-(a) = \min(0, a)$ and $M^+(a) = \max(0, a)$ for $a \in \mathbb{R}$. Clearly, $a = M^-(a) + M^+(a)$. Differentiating (1), the partial derivative of $f(x)$ with respect to x_1 is $\frac{\partial f}{\partial x_1}(x) = \sum_{i=1}^{m} a_i \frac{\partial \phi_i}{\partial x_1}(x)$. Using (4), $a_i \frac{\partial \phi_i}{\partial x_1}(x) \ge M^-(a_i) U_{i1} + M^+(a_i) L_{i1}$. Applying this result, $\frac{\partial f}{\partial x_1}(x) \ge \sum_{i=1}^{m} M^-(a_i) U_{i1} + M^+(a_i) L_{i1}$. Therefore, a sufficient condition for $f(x)$ to be non-decreasing in x_1 is $\sum_{i=1}^{m} M^-(a_i) U_{i1} + M^+(a_i) L_{i1} \ge 0$. Similarly, for $f(x)$ to be non-increasing in x_1, $\sum_{i=1}^{m} M^+(-a_i) L_{i1} + M^-(-a_i) U_{i1} \ge 0$, using the identities $-M^-(a) = M^+(-a)$ and $-M^+(a) = M^-(-a)$.

It is now possible to state sufficient conditions for monotonicity in multiple dimensions. Let

$$s_j = \begin{Bmatrix} 1 \text{ if } f \text{ shall be non-decreasing in } x_j \\ -1 \text{ if } f \text{ shall be non-increasing in } x_j \\ 0 \text{ if } f \text{ is not monotonically constrained in } x_j \end{Bmatrix}, \text{ for } j = 1, \dots, n.$$

Then the aggregate conditions for monotonicity as specified in $\{s_j\}$ are

$$\sum_{i=1}^{m} M^-(s_j a_i) U_{ij} + M^+(s_j a_i) L_{ij} \ge 0 \text{ for } j = 1, \dots, n, \quad (5)$$

where $L_{ij} \le \frac{\partial \phi_i}{\partial x_j} \le U_{ij}$, for $j = 1, \dots n,$ such that $s_j \ne 0$, and for all $x \in A$.

Next, we apply (5) to the response surface form in (3). We require the region A to be bounded by the hypercube $\prod_j [l_j, u_j]$. The partial derivatives of the terms in (3) with respect to x_1 all vanish except for $\frac{\partial}{\partial x_1} x_1 = 1$, $\frac{\partial}{\partial x_1} x_1 x_j = x_j$ for $j = 2, \dots, n$, and $\frac{\partial}{\partial x_1} x_1^2 = 2x_1$. Because of bounds on the independent variables, the partial derivatives are bounded as follows: $\frac{\partial}{\partial x_1} x_1 = 1$, $l_j \le \frac{\partial}{\partial x_1} x_1 x_j \le u_j$, and $2l_1 \le \frac{\partial}{\partial x_1} x_1^2 \le 2u_1$. Partial derivatives in the other dimensions are similarly bounded. Therefore, applying (5)

with $\{s_j\}$ again being the specification for monotonicity, the sufficient condition for monotonicity of the quadratic response surface is

$$s_j a_j + 2\left(M^-(s_j c_j)u_j + M^+(s_j c_j)l_j\right) + \sum_{\substack{i=1 \\ i\neq j}}^{n}\left(M^-(s_j b_{ij})u_i + M^+(s_j b_{ij})l_i\right) \geq 0 \quad \text{for } j = 1, \ldots n, \text{ such that } s_j \neq 0. \tag{6}$$

Condition (6) also happens to be a necessary condition for monotonicity of the quadratic response surface provided that $A = \prod_j [l_j, u_j]$. For simplicity suppose that for some choice of a_i, b_{ij} and c_i, that f in (3) is non-decreasing in x_1 for all $x \in A = \prod_j [l_j, u_j]$. In particular, choose x as follows: if $c_1 < 0$, then $x_1 = u_1$, otherwise $x_1 = l_1$; if $b_{i1} < 0$ then $x_i = u_i$, otherwise $x_i = l_i$. Then for this x,

$$\begin{aligned} 0 \leq \frac{\partial f}{\partial x_1}(x) &= a_1 + \textstyle\sum_{i=2}^{n} b_{i1}x_i + 2c_1x_1 \\ &= a_1 + \sum_{i=2}^{n}(M^-(b_{i1})u_i + M^+(b_{i1})l_i) + 2(M^-(c_1)u_1 + M^+(c_1)l_1), \end{aligned}$$

which is (6) with $j = 1$ and $s_1 = 1$. Similarly, if f is non-increasing in x_1, we can choose $x \in A = \prod_j [l_j, u_j]$ such that

$$0 \leq -\frac{\partial f}{\partial x_1}(x) = -a_1 + \textstyle\sum_{i=2}^{n}(M^-(-b_{i1})u_i + M^+(-b_{i1})l_i) + 2(M^-(-c_1)u_1 + M^+(-c_1)l_1).$$

In general, (6) is true if f is monotonic as specified by $\{s_j\}$.

3 STATISTICAL ANALYSIS WITH MONOTONIC MODELS

This section lays out a statistical model based on the monotonic response surfaces (1) and (3). A model for N simulation responses is

$$y_k = f(x_k) + \varepsilon_k, \text{ for } k = 1, \ldots, N, \tag{7}$$

where f is one of the forms in (1) or (3), $x_k \in A \subset \mathbb{R}^n$ are design points, y_k is the simulation response for design point x_k and ε_k is error due to random sampling in the simulation. We will assume that ε_k are independent and normally distributed, $\varepsilon_k \sim N(0, \sigma^2)$; σ^2 is the common variance for $\{\varepsilon_k\}$.

The log likelihood function for the sample in (7) with f from (1) is

$$\mathcal{L}(a, \sigma | y) = -\frac{1}{2}N\log(2\pi) - N\log(\sigma) - \frac{1}{2}\sum_{k=1}^{N}\frac{\left(y_k - \sum_{i=1}^{m} a_i\phi_i(x)\right)^2}{\sigma^2}. \tag{8}$$

with $y = (y_1, \ldots, y_N)$ and $a = (a_1, \ldots, a_m)$.

Then the maximum likelihood monotonic response surface depends on the estimates $\hat{a}, \hat{\sigma}^2$ which maximize (8) subject to constraints (5). One may solve this problem using a constrained nonlinear optimization utility, such as MATLAB's fmincon. See Section 5 for an example.

More general models of sampling error, such as non-normal distributions, non-uniform variances and correlation between errors are possible. See for example, Gelfand (1997), Lim and Glynn (2006) and Staum (2009). These different models will lead to different forms for the log likelihood. Nevertheless, one may pursue the same approach, which is to maximize the likelihood function subject to constraints on the parameters which will ensure monotonicity. In the case of GLM, note that link functions preserve monotonicity. The method of constraining linear model coefficients enforces monotonicity as well with GLM as in simple linear models.

One should make sure that the simulation response really is monotonic before imposing monotonic constraints on the response surface model. Simulations often produce non-intuitive non-monotonic results

as a result of bugs in implementation, invalid simulation models of processes, or a modeled process that is not in truth monotonic. In the two former cases, the simulation and models need more scrutiny leading to debugging or fidelity improvements. The latter case could lead to an unexpected and important finding.

4 BAYESIAN MONOTONIC RESPONSE SURFACES

A Bayesian inference of the model in (7) with response surface defined by (1) or (3) needs only prior distributions on the parameters $\{a_i\}$ and σ^2 together with a means of estimating the posterior distribution, or at least a means of simulating a sample from the posterior. In this paper, we use bounded uniform prior distributions.

Let the prior density of $\{a_i\}$ be uniform on the region of support defined by $-\Omega < a_i < \Omega$, together with the constraints (5), where Ω is sufficiently large to bound all plausible values of a_i. Let the density of σ be uniform on $\epsilon \leq \sigma \leq \Omega$ with ϵ sufficiently small and Ω set sufficiently large to bound all plausible values for σ. These assumptions give the parameters non-informative proper prior densities. The log joint density of a, σ, y, ignoring the normalization constant is

$$\mathcal{L}(a,\sigma,y) = -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2}\sum_{k=1}^{N}\frac{\left(y_k - \sum_{i=1}^{m} a_i\phi_i(x)\right)^2}{\sigma^2}, \tag{9}$$

with $\epsilon \leq \sigma \leq \Omega$, $-\Omega < a_i < \Omega$ and a_i subject to (5). The density (9) represents a proper joint density.

In absence of a closed-form expression for the posterior density $\mathcal{L}(a,\sigma|y)$, Monte Carlo Markov Chains (MCMC) provide a family of algorithms for sampling from $\mathcal{L}(a,\sigma|y)$ without integration (Chib and Greenberg 1995). All of the numerical examples in this paper were evaluated using the Differential Evolution Markov Chain from Ter Braak (2005).

Instead of the hard constraints of (5), we multiply the joint density by steep roll-off functions of the form $\mathcal{S}(a) = exp\left(-\left(\frac{M^-(a)^2}{\epsilon^3}\right)\right)$, where ϵ is small enough that $\mathcal{S}(a)$ approximates a zero-centered step function. Specifically, ϵ is sufficiently small that the loss from $\mathcal{S}(a)$ is much greater than losses from lack of fit expressed in (9). The log joint density when modified by these multipliers becomes

$$\begin{aligned}\mathcal{L}'(a,\sigma,y) = &-\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2}\sum_{k=1}^{N}\frac{\left(y_k - \sum_{i=1}^{m} a_i\phi_i(x)\right)^2}{\sigma^2} \\ &+ \sum_{j=1}^{n}\log\left(\mathcal{S}\left(M^-(s_j a_i)U_{ij} + M^+(s_j a_i)L_{ij}\right)\right).\end{aligned} \tag{10}$$

Unlike the ordinary step function, $\log\mathcal{S}(a)$ is defined for $a < 0$. When the MCMC proposes samples that violate constraints, (10) imposes a stiff penalty. Markov chains migrate onto the support of the prior distributions.

5 NUMERICAL EXAMPLES OF SIMPLE MONOTONIC RESPONSE SURFACES

Returning to our example from the introduction, what would happen if the first-order model of the response (2) were re-estimated applying prior knowledge that the function is non-decreasing in factors A, B and C? Equation (6) reduces to

$$\begin{aligned} a_A + \left(M^-(b_{AB}) - M^+(b_{AB})\right) + \left(M^-(b_{AC}) - M^+(b_{AC})\right) + \left(M^-(b_{AD}) - M^+(b_{AD})\right) \geq 0 \\ a_B + \left(M^-(b_{AB}) - M^+(b_{AB})\right) + \left(M^-(b_{BC}) - M^+(b_{BC})\right) + \left(M^-(b_{BD}) - M^+(b_{BD})\right) \geq 0 \\ a_C + \left(M^-(b_{AC}) - M^+(b_{AC})\right) + \left(M^-(b_{BC}) - M^+(b_{BC})\right) + \left(M^-(b_{CD}) - M^+(b_{CD})\right) \geq 0, \end{aligned} \tag{11}$$

since there are no second-order effects making $c_j = 0$ for all j, and because the variables are from $\{-1, 1\}$, $l_i = -1$ and $u_i = 1$ for all i.

A new maximum likelihood estimate based on (8) and constrained by (11) is

$$\begin{aligned}\hat{\hat{f}}(A,B,C,D,E) = 1.551 + (0.954)A + (0.330)B + (0.149)C - (0.028)D + (0.077)E - \\ (0.185)AB + (0.000)AC + (0.081)AD - (0.071)BC + (0.012)BD + (0.071)CD.\end{aligned} \tag{12}$$

The constrained coefficients were computed using MATLAB's fmincon function. Figure 2 shows predictions of the constrained model for $A = E = -1$.

The constrained response surface preserves monotonicity for the three variables known to possess it. When compared to the true function over the full factorial space of 32 inputs, the constrained model's root mean squared error (RMSE) was 0.25 compared with the unconstrained model's RMSE of 0.29. In this case, a constrained response surface estimator complies with prior knowledge of monotonicity with no sacrifice in prediction error.

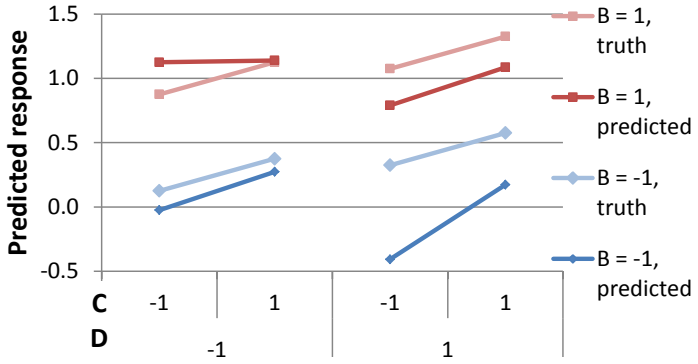

Figure 2: Predicted response from constrained regression model (12) of response data (Table 1).

A more practical application comes from simulation of integrated air and missile defense (IAMD) systems. Table 2 shows 12 parameters that control a scenario in which multiple cruise missiles attack a high value site defended by an IAMD. The IAMD consists of radars, a command and control net, missile interceptors and launchers. The parameters are normalized to the interval [0,1] using their minimum and maximum values. Table 2 distinguishes between uncertainty parameters (those dependent on an uncertain threat and environment) and design parameters (attributes of the IAMD). In this application, IAMD developers explore responses to different combinations of parameters in order to decide on investments in IAMD improvements.

Table 2: Parameters governing simulation of site defense against cruise missiles.

Factor	Type	Minimum	Maximum
Terrain roughness	Uncertainty	0: flat	1: hilly
Number of threats	Uncertainty	0: two cruise missiles	1: 12 cruise missiles
Threat spacing	Uncertainty	0: small	1: large
Single shot PK	Design	0: low	1: high
Weapon range	Design	0: short	1: long
Weapon speed	Design	0: slow	1: fast
Salvo size	Design	0: one per threat	1: two per threat
Launcher loadout	Design	0: small	1: large
Simultaneous engagements	Design	0: one per launcher	1: many per launcher
Reaction time	Design	0: short	1: long
Surveillance radar range	Design	0: short	1: long
Fire control radar range	Design	0: short	1: long

A Monte Carlo simulation outputs the "number of leakers", which is the number of cruise missiles in the scenario not defeated by the IAMD. The simulation was executed for a Latin hypercube DOE of 120 unique cases with 10 Monte Carlo repetitions per case, resulting in 1,200 measurements of leakers. The mean of leakers for the 120 cases ranged from zero to 6.6 with an overall average of 0.96 leakers. Using

the 10-per-case Monte Carlo repetitions, the standard errors of the means ranged from zero to 0.60 with an overall root mean square of 0.22.

For comparison, we fit two types of response surfaces to this data set, stochastic kriging (Staum 2009) and a monotonically constrained GLM.

The stochastic kriging response surface was a superposition of a 13-term linear model (first-order trends without interactions) and a Gaussian process with covariance Σ such that $\Sigma_{ij} = \tau^2 \exp\left(-\sum_{k=1}^{d} \rho_k \left(X_{ik} - X_{jk}\right)^2\right)$. The 13 linear coefficients, τ^2, and 12 values ρ_k were chosen to maximize the likelihood function assuming normal sampling error.

The stochastic kriging response surface fits the data set of 120 mean leakers with an RMSE of 0.2 leakers. However, in a k-fold cross validation (Arlot 2010) with k = 10, the RMSE of the kriging surfaces was 1.0 leakers. Stochastic kriging virtually always fits to within the standard error of the mean at the DOE design points. In between points, it smoothly reverts to the linear model (Staum 2009). In experiments with very many dimensions, the d-dimensional space between points is quite large, resulting in cross validation errors much larger than errors of fit.

Figure 3 contains profiles of the stochastic kriging response surface around an arbitrary center point indicated by the vertical lines. Of particular interest is the incorrect non-monotonic response to reaction time. Likewise, the weapon range response in Figure 3 is counter-intuitive. Leakers should increase with increasing reaction time and decrease with increasing weapon range in the IAMD simulation. The non-monotonic trends in Figure 3 are a result of sampling error and an unconstrained model space.

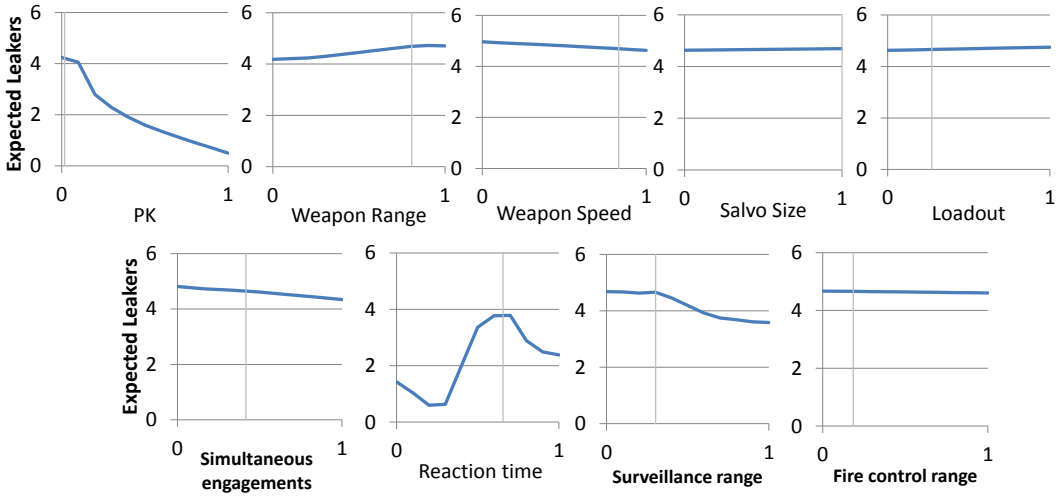

Figure 3: Stochastic kriging profile of nine factors in the IAMD site defense simulation around a center point: terrain: 1, number of threats: 0.30, threat spacing: 0.388, probability of kill: 0.02, weapon range: 0.81, weapon speed: 0.83, salvo size: 1, launcher loadout: 0.27, simultaneous engagements: 0.41, reaction time: 0.66, surveillance radar range: 0.30, fire control radar range: 0.19.

For a Bayesian monotonically constrained response surface for the IAMD site defense data set, we must first identify those factors known to be monotonic and the direction of monotonicity for each such factor. Cruise missile leakers is non-decreasing with respect to three factors: terrain, number of threats and reaction time. Leakers is non-increasing with respect to eight factors: probability of kill, threat spacing, weapon range, weapon speed, launcher loadout, simultaneous engagements, surveillance radar range and fire control radar range. For the one factor, salvo size, the direction of effect is unknown or non-monotonic. Instead of the linear mean response model in (1), we use a GLM of the form

$$y = [(y_{max}(x) + 1)g(f(x) + \varepsilon)], \tag{13}$$

where $g(x) = \left(1 + exp(-x)\right)^{-1}$ is the logistic link function, $y_{max}(x)$ is the number of threats for design point x and $[y]$ denotes the largest integer less than y. The sampling errors ε are independent and normal, with zero mean and uniform variance. In short, y is a scaled and discretized logit normal distribution. It is discrete and bounded to the same range as the leakers output of the IAMD simulation. We use form (3) for $f(x)$. $c_i = 0$ for the factors terrain and salvo size, because these factors have only two levels.

The parameters to be estimated by a sample from their posterior distributions are:

- $a_i, b_{ij},$ and c_i, where i, j are one of the 12 factors, with the exception $c_i = 0$ for the factors terrain and salvo size,
- σ, the standard deviation of the sampling error ε_{kl} for Monte Carlo replication l of unique case k. ε_{lk} is the same as ε in (13).

Using this measurement model for y and the non-informing priors assumed in Section 4, the posterior probability density of $a_i, b_{ij}, c_i,$ and σ given the 1,200 measurements of leakers y_{kl} is

$$p(a,b,c,\sigma|y) \propto \prod_{k=1}^{120}\prod_{l=1}^{10}\left(\Phi\left(\left(g^{-1}\left(\frac{y_{kl}+1}{y_{max(x_k)+1}}\right) - f(x_k)\right)\sigma^{-1}\right) - \Phi\left(\left(g^{-1}\left(\frac{y_{kl}}{y_{max(x_k)+1}}\right) - f(x_k)\right)\sigma^{-1}\right)\right)$$

on the support of a_i, b_{ij}, c_i, σ, which includes monotonicity constraints (6). Φ is the standard normal cumulative distribution function, $\Phi\left(g^{-1}(y)\right)$ is defined to be zero if $y \le 0$ and one if $y \ge 1$. The MCMC algorithm resulted in a a sample of 100 instances of $a_i, b_{ij}, c_i,$ and σ, namely $\left\{ a_{il}, b_{ijl}, c_{il}, \sigma_l;\ i,j = 1,..,n;\ l = 1,..,100\right\}$. A Monte Carlo prediction of the mean response is

$$\hat{y} = \frac{1}{100}\sum_{l=1}^{100}\left[y_{max}(x)\, g\left(a_{0l} + \sum_{i=1}^{n} a_{il}x_i + \sum_{i=1}^{n}\sum_{j=i+1}^{n} b_{ijl}\, x_i x_j + \sum_{i=1}^{n} c_{il}x_j^2 + \varepsilon_l\right)\right]$$

where ε_l is an additional Monte Carlo sample from $N(0, \sigma_l)$. The purpose of simulating sampling error is to include the bias in averaging over the nonlinear function g.

This model has an RMSE of 0.9 leakers in cross-validation. Table 3 compares the fit and cross-validation errors of the kriging and Bayes MCMC predictors. As expected, the monotonic model did not fit the simulation responses as well as kriging. However, it performed as well as kriging in cross-validation.

Table 3: Prediction error summary for stochastic kriging and monotonic quadratic models applied to the IAMD site defense simulation response.

Predictor	RMSE of fit to 120 responses	RMSE of 10-fold cross-validation
Stochastic kriging	0.2	1.0
Monotonic quadratic	0.9	0.9

Figure 4 is a profile of the monotonic model at the same center point as Figure 3. Like the kriging estimator, the monotonic estimator identifies PK and reaction time as the two strongest trends. However, unlike kriging, the monotonic estimator is uniformly increasing with respect to reaction time. Also, while kriging showed a trend reversal with respect to weapon range, the monotonic model has a nearly flat

response. The monotonic model shows as good predictive capability as kriging without showing non-intuitive trends.

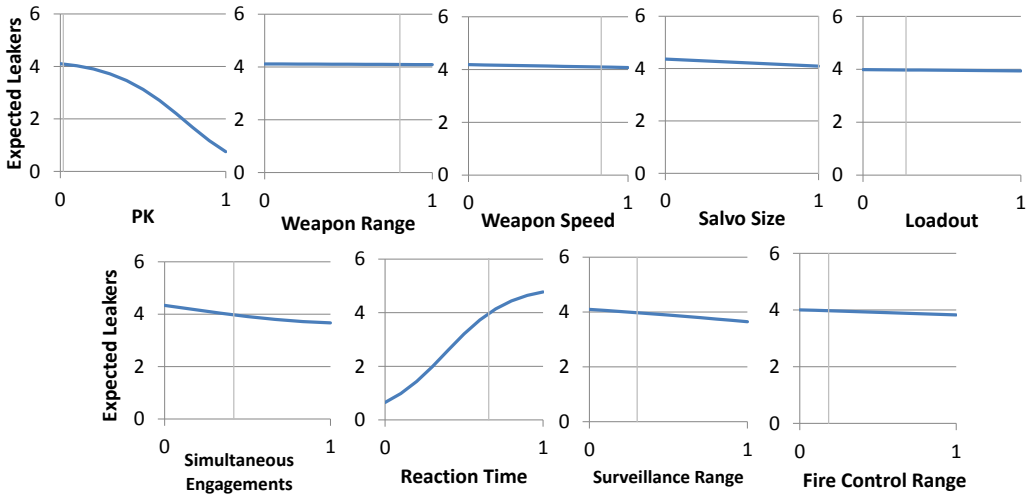

Figure 4: Monotonic quadratic response surface profile of nine factors in the IAMD site defense simulation around the center point: terrain: 1, number of threats: 0.30, threat spacing: 0.388, probability of kill: 0.02, weapon range: 0.81, weapon speed: 0.83, salvo size: 1, launcher loadout: 0.27, simultaneous engagements: 0.41, reaction time: 0.66, surveillance radar range: 0.30, fire control radar range: 0.19.

6 SUMMARY AND DISCUSSION

Quadratic response surfaces are accessible to a large population of simulation analysts. We offer an option to constrain quadratic responses to produce suitable monotonic estimates. Monotonic quadratic linear models produce intuitive well-behaved estimates in cases with many monotonic factors and austere designs of experiments.

Numerical estimation of monotonic response surfaces is feasible either by constrained maximum likelihood or Bayes MCMC methods. We here demonstrated Bayes estimation with up to 12 factors, 120 unique cases, and 1,200 simulations. The method adapts well to non-linear link functions and sampling error models that generate bounded and discrete responses.

We showed that the monotonic conditions (6) are both necessary and sufficient for quadratic response surfaces. However, necessity depends on an assumption that all points in the hypercube are valid inputs. For some applications, this may not be so. In such cases, the monotonic conditions may be too restrictive.

We limited the statistical applications and examples to cases with independent sampling errors. This is not the case for many simulation applications.

We restricted our example to GLMs with a fixed link function. As noted in Section 1, Dirichlet mixture generators provide a more general class of link functions (Mallick and Gelfand 1994, Gelfand 1997). We might consider combining a generalized link function model with monotonic constraints on the linear coefficients.

Computations with the IAMD application revealed that there is much room for algorithmic improvements. The Bayes MCMC implementation did not settle nearly as fast in the IAMD site defense case as it has in unconstrained model fitting and in lower-dimensional problems. It helped to mix the constrained maximum likelihood estimator with the MCMC streams on initialization to help point to the main mode of the posterior distribution.

REFERENCES

Arlot, S., Celisse, A. 2010. "A Survey of Cross-validation Procedures for Model Selection." *Statistics Surveys* 4: 40-79.

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., Silverman, E. 1955. "An Empirical Distribution Function for Sampling with Incomplete Information." *Annals of Mathematical Statistics* 26: 641-647.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. , Brunk, H. D. 1972. *Statistical Inference Under Order Restrictions*. London: Wiley, Inc.

Burdakow, O., Grimwall, A., Hussain, M. 2004 "A Generalized PAV Algorithm for Monotonic Regression in Several Variables." *COMPSTAT 2004 Symposium*, edited by J. Antoch, Heidelberg, 761-767. NY: PhysicaVerlag/Springer.

Chib, S. and Greenberg, E. 1995. "Understanding the Metropolis-Hastings Algorithm." *American Statistician* 49: 327-335.

Gelfand, A. E. 1997 "Approaches for Semiparametric Bayesian Regression." *Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models, Journal of Computational and Graphical Statistics*, pp. 615-638, Marcel Dekker, Inc.

Gelfand, A. E. and Kuo, L. 1991 "Nonparametric Bayesian Bioassay Including Ordered Polytomous Response." *Biometrika* 78: 657-666.

Hall, P. and Huang, L. (2001) "Nonparametric Kernel Regression Subject to Monotonicity Constraints", *Annals of Statistics*, 29(3) pp. 624-647.

Kay, H. and Ungar, L. 2000 "Estimating Monotonic Functions and Their Bounds." *AIChE Journal* 46: 2426-2434.

Leoppky, J. L. , Sacks, J., Welsh, W. 2009 "Choosing the Sample Size of a Computer Experiment: A Practical Guide." *Technometrics* 51: 366-376.

Lim, E. and Glynn, P. W. 2006 "Simulation-Based Response Surface Computation in the Presence of Monotonicity." *Proceedings of the 2006 Winter Simulation Conference,* edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto. 264-271. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lin, L. and Dunson, D. (2014) "Bayesian Monotone Regression Using Gaussian Process Projection", *Biometrika*, 101(2) pp. 303-317.

Mallick, B. K. and Gelfand, A. E. (1994) "Generalized Linear Models with Unknown Link Functions", *Biometrika*, 81(2) pp. 237-245.

Neelon, B. and Duson, D. B., (2004) "Bayesian Isotonic Regression and Trend Analysis", *Biometrika*, 60(2), pp. 398-406.

NIST/SEMATECH 2012 "How Do You Select an Experimental Design?" *e-Handbook of Statistical Methods*, Accessed 29 June 2014 from http://www.itl.nist.gov/div898/handbook/pri/section3/pri33.htm.

Penn State 2014 "Lesson 11: Response Surface Methods and Designs." *STAT 503 Design of Experiments*, Accessed 29 June 2014, https://onlinecourses.science.psu.edu/stat503/node/57.

Racine, J. S. and Parmeter, C. F. 2008 "Constrained Nonparametric Kernel Regression: Estimation and Inference." Maxwell School of Citizenship and Public Affairs, Accessed 29 June 2014 from https://www.maxwell.syr.edu/uploadedFiles/econ/kernel_cons.pdf?n=6322

ReliaSoft Corporation 2013 "Chapter 10: Response Surface Methods for Optimization." *Experiment Design and Analysis Reference*. Accessed 29 June 2014 from http://reliawiki.org/index.php/Response_Surface_Methods_for_Optimization.

Riihimäki, J. and Vehtari, A. 2010 "Gaussian Processes with Monotonicity Information." *13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, edited by N. Lawrence and N. Reid, Journal of Machine Learning Workshop and Conference Proceedings. 645-652.

SAS Institute 2014 "Response Surface Designs." *JMP 11 Online Documentation*. Accessed 29 June 2014 from http://www.jmp.com/support/help/Response_Surface_Designs.shtml.

Staum, J. 2009 "Better simulation Metamodeling: the How, What and Why of Stochastic Kriging." *Proceedings of the 2009 Winter Simulation Conference,* edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc. 119-133.

Ter Braak, C. J. F. 2005 "A Markov Chain Monte Carlo Version of the Genetic Algorithm Differential Evolution: Easy Bayesian Computing for Real Parameter Spaces." *Statistics and Computing* 16:239-249.

Xu, H. 2013 "Chapter 11. Response Surface Methods and Designs." Presentation, *Stats 201A: Research Design, Sampling and Analysis*. Accessed 20 June 2014 from http://www.stat.ucla.edu/~hqxu/stat201A/ch11.pdf.

AUTHOR BIOGRAPHY

FREDERICK A. AHRENS is a Senior Principal Systems Engineer in the Operations Research Department, Raytheon Missile Systems, practicing modeling and simulation of operational utility of force protection and civil security capabilities for over 30 years, using probabilistic risk analysis, CONOPS development, designs of experiments, response surface analysis, and visual analytics. He holds a B.S. in Mathematics from Harvey Mudd College, an M.A. in Statistics from the University of California at Berkeley, and a M.S. in Systems Management from University Southern California. His email address is faahrens@raytheon.com.