# SANDIA REPORT

SAND2013-5145
Unlimited Release
Printed June 2013

# New Methods of Uncertainty Quantification for Mixed Discrete-Continuous Variable Models

Lara Bauman

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.



Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.



Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2013-5145
Unlimited Release
Printed June 2013

# New Methods of Uncertainty Quantification for Mixed Discrete-Continuous Variable Models

Lara E. Bauman
Quantitative Modeling & Analysis
lebauma@sandia.gov
Sandia National Laboratories
PO Box 969, Mail Stop 9159
Livermore, CA 94551-0969

**Abstract**

The scale and complexity of problems such as designing power grids or planning for climate change is growing rapidly, driving the development of complicated computer models. More complex models have longer run times and incorporate larger numbers of inputs, both continuous and discrete. For example, a detailed physics model may have continuous variables such as temperature, height or pressure along with discrete variables that indicate the choice of a material for a particular piece or the model to be used to calculate air flow. A power grid design model may have continuous variables such as generation capacity, power flow or demand along with discrete varibles such as number of generators, number of transmission lines or binary variables to indicate whether or not a node is chosen for generation expansion [4]. A growing awareness of uncertainty and the desire to make risk-informed decisions is causing uncertainty quantification (UQ) to be more routine and often required. UQ provides the underpinnings necessary to establish confidence in models and their use; therefore, much time and effort is being invested in creating efficient approaches for UQ. However, these efforts have been focused on models that take continuous variables as inputs. When discrete inputs are thrown into the mix, the basic approach is to repeat the UQ analysis for each combination of discrete inputs or some subset thereof; this rapidly becomes intractable. Because of the computational complexity inherent in mixed discrete-continuous models, researchers will focus on the uncertainty in their particular problem finding ways to take advantage of symmetries, simplifications or structures. For example, uncertainty propagation in certain dynamical systems can be efficiently carried out after various decomposition steps or uncertainty propagation in stochastic programming is confined to scenario generation. Unfortunately models are not always available for such machinations: models may be embedded in legacy codes, may utilize commercial off the shelf codes or may be created by stringing a series of codes together. It is also time consuming to start each problem from scratch; worse there may not be any simplifications or symmetries to take advantage of. For these situations a UQ method developed for any black box function is necessary. This report documents a new conceptual model for performing UQ for mixed discrete-continuous models which not only applies to any simulator function, but allows the use of the efficient UQ methods that have been developed for continuous inputs only. The conceptual model is presented and an estimation procedure is fleshed out for one class of problems. This is applied to variations of a mixed discrete-continuous optimization test problem. This procedure provides comparable results to a benchmark solution with fewer function evaluations.

# Acknowledgment

I would like to acknowledge the tremendous help of Genetha Gray, Patricia Hough and Jerry McNeish. I would like to thank the LDRD office and the Early Career program for providing funding for this work.

This page intentionally left blank.

# Contents

# List of Figures

This page intentionally left blank.

# Chapter 1

# Introduction

The scale and complexity of problems such as designing power grids or planning for climate change is growing rapidly, driving the development of complicated computer models. More complex models have longer run times and incorporate larger numbers of inputs, both continuous and discrete. For example, a detailed physics model may have continuous variables such as temperature, height or pressure along with discrete variables that indicate the choice of a material for a particular component or the model to be used to calculate air flow. A power grid design model may have continuous variables such as current, generation capacity, power flow or demand along with discrete varibles such as number of generators [4]. A growing awareness of uncertainty and the desire to make risk-informed decisions is making uncertainty quantification (UQ) more routine. UQ provides the underpinnings necessary to establish confidence in models and their use; therefore, much time and effort is being invested in creating efficient approaches for UQ. However, these efforts have been focused on models that take continuous variables as inputs. When discrete inputs are thrown into the mix, the basic approach is to repeat the UQ analysis for each combination of discrete inputs or some subset thereof; this rapidly becomes intractable. This report documents a new conceptual model for performing UQ for mixed discrete-continuous models which not only applies to any simulator function, but allows the use of the efficient UQ methods that have been developed for continuous inputs only. The conceptual model and estimation procedure is described for one class of problems and applied to a mixed discrete-continuous optimization test problem. This procedure provides comparable results to a benchmark solution with fewer function evaluations and opens up a new conceptual model for UQ on discrete-continuous models.

Because of the computational complexity inherent in mixed discrete-continuous models, first researchers will try to focus on the uncertainty in their particular problem finding ways to take advantage of symmetries, simplifications or structures. For example, uncertainty propagation in certain dynamical systems can be efficiently carried out after various decomposition steps or uncertainty propagation in stochastic programming is confined to scenario generation. Unfortunately models are not always available for or amenable to such machinations, for example, models may be embedded in legacy codes or there may not be a straightforward way to simplify the problem. In this case, the second step typically relies on one of three concepts in order to take advantage of the more efficient UQ methods designed for continous inputs:

1. Separate analysis

Perform a separate analysis for each unique combination of discrete variables which means any UQ method designed for continuous inputs can be employed. While this is the most comprehensive analysis, it becomes computationally intensive for even modest numbers of discrete variables. It is often unclear how to combine the results from each separate analysis into one global uncertainty measurement.

2. Choose one discrete configuration

   Assume isotropic model behavior across all unique combinations of discrete variables. While it is unrealistic to assume that behaviors or correlations calculated with one combination of discrete variables will extend to all other combinations, UQ methods designed for continuous inputs can still be used.

3. Treat a discrete variable as if it were a continuous one.

   This is unsatisfactory at best, e.g., for a variable such as height of an air vent unit location (low, medium or high), and incomprehensible at worst, e.g., when a component can be made of one of four materials. Once again, UQ methods designed for continuous inputs can be used.

The basic UQ process has its roots in the law of large numbers. It is based on sampling the model or some approximation to the model [13, 28]. Summary statistics are then calculated from the outputs such as the mean, variance or pdf. The output pdfs are discrete approximations to the "true" pdf of model output values. The efforts expended to create more efficient UQ methods for continuous inputs are focused on three things:

A. sampling procedures

   More efficient sampling procedures strive to get the maximum information on which to base the output statistics for the minimum number of actual functional evaluations. The statistics here are calculated using actual function evaluations. Latin hypercube sampling (LHS) [12], importance sampling [26] or grids are examples of efficient sampling approaches.

B. reliability methods

   Reliability methods seek to answer the question: what is the probability the function value will exceed $k$? *expand a bit here?*

C. surrogate models or response surfaces

   These are approximating surfaces based on a small set of model evaluations. Samples from this approximate surface provide the base for the output statistics. Gaussian process models (GP), spline fits or polynomial chaos expansions (PCE) [8] are examples of surrogate approaches.

None of these methods were developed with discrete inputs in mind; in fact, many rely on smoothness or continuity assumptions that make direct application suspect or impossible. LHS is the only method that easily generalizes to both discrete and and continuous inputs.

This report introduces a new concept: aggregating and transforming discrete variables into continuous probabilistic ones. This concept provides a natural means of performing uncertainty quantification on mixed variable models that is significantly more tractable than doing separate analyses on each combination of discrete variables (as in 1 above) while retaining more information than using one set of discrete combinations (as in 2 above) or treating a discrete variable as if it were continuous (as in 3).

The key idea is to replace a set of discrete variables by an associated probability, for example, the probability the set of variables has a particular effect on a quantity of interest. This approach has two distinct advantages: one, it directly incorporates uncertainty into the model by taking into account that the state or effect of a discrete variable may be unknown, and two, it makes possible the use of efficient UQ methods designed for continuous inputs. This approach is not related to optimization nor is it sensitivity analysis, but an efficient way to determine the overall variability in output values given the uncertain continuous inputs and the variability stemming from the complete set or some subset of discrete variables. This approach has one more desirable property: it satisfies the simplicity property laid out by [22], "sets or subgroups of inputs can be treated as single entities (factors)".

Detailed analysis is required to delineate classes of problems for which this new approach will be effective, especially since the nature of the probabilistic transformation will be different for each class of problems. In this report I define a class of problems called model choice problems; a second possible class of problems is presented in the discussion. Model choice problems have a moderate number of discrete variables plus continuous variables and moderate to long simulator run times. These characteristics make full LHS extremely time consuming and the moderate number of discrete variables prevents a full analysis for each separate configuration of discrete variables. For example, engineering applications are detailed physics-based models with a variety of continuous and discrete inputs. Examples of discrete inputs include a metal/material for a specific part or the model used to calculate turbulence. Many of the discrete choices will be associated with continuous properties within the black box of the simulator; for example, a material choice will have a hardening curve, density or melting temperature. The both the discrete and continuous variables are assumed to have known or estimated probability distributions. The model output (also called function, simulator, objective function, cost function) is known and continuous with a single output; extension to multiple outputs is straightforward.

This page intentionally left blank.

# Chapter 2

# Methods

## 2.1 Conceptual Model

The key to this new approach is to *transform* the discrete inputs into a continuous inputs so that UQ methods designed for continuous inputs can be used. One can also think of this process as summarizing the net effect of the discrete choices or as averaging over the discrete effects. Let $f$ represent the output function of discrete inputs $D$ and continuous inputs $C$. The traditional UQ process is:

1. define the densities for all inputs
2. propagate the uncertainties through the function

Figures 2.1 to 2.3 illustrate the new concept which has three general steps:

1. decompose the function $f(D, C)$ (See Figure 2.1)
2. transform the discrete variable into a continuous variable and estimate the empirical probability distribution function of this continuous variable (epdf) of the effects of the discrete variables (See Figure 2.2)
3. use a UQ method designed for continuous inputs to propagate the uncertainties through the function (See Figure 2.3)

The transformation of the discrete variables into a continuous random variable starts by forcing an additive decomposition of the original function.

$$f(D, C) \approx g_D(D) + g_C(C) \tag{2.1}$$
$$\approx X + g_C(C) \tag{2.2}$$

where the random variable $X$ has a probability density function equal to the pdf of $g_D(D)$. The word "force" is chosen to emphasize the fact that such a decomposition may not be additive in actuality. Because of this, any function with interactions between its discrete and continuous variables will always have an approximate solution, never an exact one; see the discussion for ideas on how to capture the error. With this decomposition in hand, the pdf of

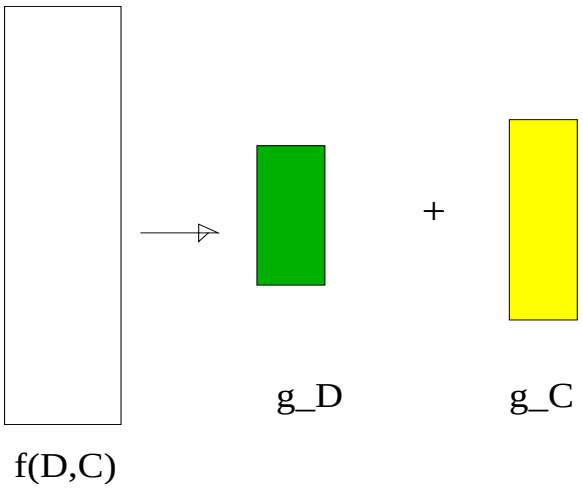



**Figure 2.1.** UQ process step one: decompose $f(D, C)$ into $g_D$ and $g_C$

$X = g_D$ can be summarized with a histogram or estimated. Since the effects of the discrete variables are now summarized in the pdf of a continuous variable, any UQ method designed for continuous inputs can be employed to characterize the uncertainty in the entire problem.

The biggest hurdle is to efficiently estimate the functions $g_D$ and $g_C$ which are generally high dimensional functions. The estimated probability distribution of $X = g_D$ must capture the net effect of all discrete variables including any correlated effects induced by combinations of discrete variables. Furthermore, it must accurately represent functional outputs. The estimated probability distribution of $X = g_D$ describes the variability in the system due to changes in discrete variables and is subsequently employed in the full uncertainty analysis. As a side note, the distribution of values for each variable set can be retained to give some information about variability sensitivities. After the effects of the discrete variables are captured, the uncertainty analysis is carried out with a UQ method designed for continuous inputs.

## 2.2 Functional Decomposition

Functional decompositions are common techniques for numerical investigations of complex functions, i.e., linear regression, Taylor series, linearization. Sobol' defined an ANOVA-representation in [24] for an integrable function $f(\mathbf{x})$ where $\mathbf{x}$ has dimension $k$ as:

$$f_D(\mathbf{x}) = f_0 + \sum_{i}^{k} f_{xi}(x_i) + \sum_{i<j}^{k} f_{xixj}(x_i, x_j) + \ldots + g_{x1x2\ldots xk}(x_1, x_2, \ldots, x_k)$$

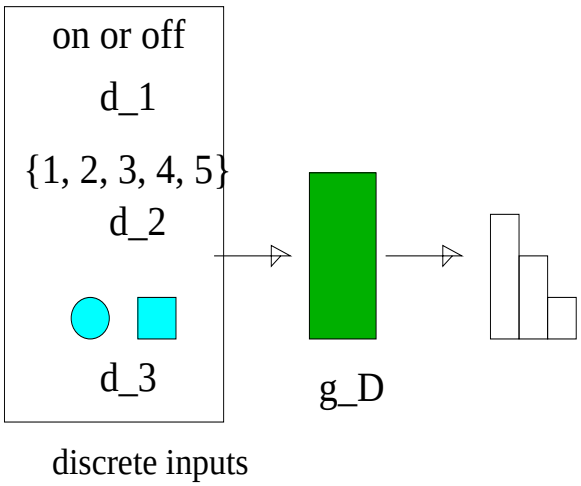



**Figure 2.2.** UQ process step two: estimate epdf of $g_D$

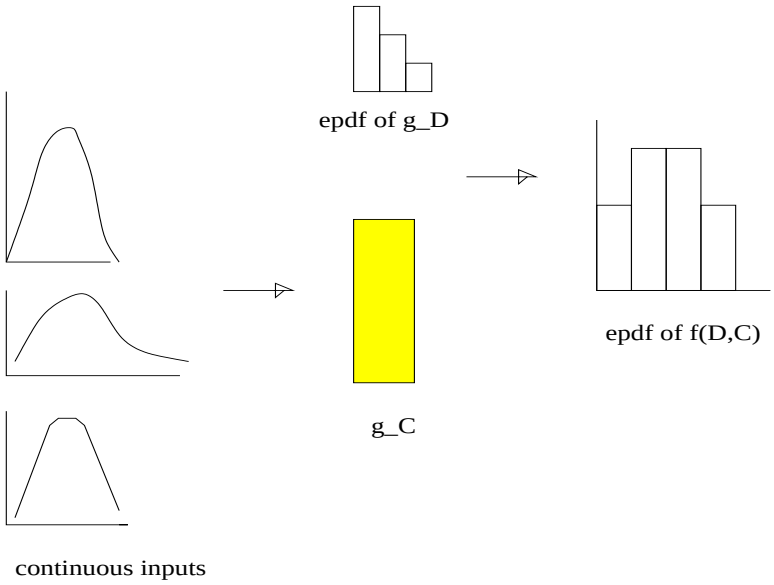



**Figure 2.3.** UQ process step three: use a UQ method designed for continuous inputs to estimate the epdf of $f(D, C)$

(integrable and square integrable functions are defined in the appendix). (define component functions) Rabitz et al. use the ANOVA-representation as the basis for high dimensional model representation (HDMR) asserting that lower dimensional component functions are sufficient to represent most functions due to low levels of input interactions. A low level of input interactions means that most inputs in their interactions with large numbers of other variables simply do not appreciably impact the output function [15]. Equation 2.1 can be viewed as a type of ANOVA/HDMR representation where interactions between discrete and continuous variables are suppressed, but interactions within discrete or continuous variable sets are allowed. Note that the ANOVA/HDMR decomposition gives no guidance on how to estimate the component functions. Integration could be used to project the original function down to the component function's dimensions, but this would involve many more functional evaluations than any direct UQ method. This is an area of active research; see for example [17, 16, 5, 14, 10].

## 2.3 Model Choice Problems

Recall that model choice problems involve moderate numbers of discrete variables and long run times and that the new concept requires the probabilistic effect of all possible discrete choices on the model to be captured within a single continuous variable $X = g_D$ which can then be used in a UQ method designed for continuous inputs (see Figure 2.1).

### 2.3.1 The form of $g_D$ using ANOVA/HDMR

The ANOVA/HDMR decomposition of $g_D$ where $\mathbf{D} = (d_1, d_2, \ldots, d_k)$ are the $k$ discrete inputs, and $\mathbf{C} = (c_1, c_2, \ldots, c_l)$ are the $l$ continuous inputs, takes this form:

$$g_D(\mathbf{D}) = f_0 + \sum_{i}^{k} g_{di}(d_i) + \sum_{i<j}^{k} g_{didj}(d_i, d_j) + \ldots + g_{d1d2\ldots dk}(d_1, d_2, \ldots, d_k).$$

The first term, $f_0$, is the average value of the original function. The first summation is over functions of single variables, the next over pairs of variables, the next over triples, etc. The last term contains any residual contributions from interactions of all the discrete inputs. The decomposition in its entirety will be equal to the original function and unique given certain orthogonality conditions on the $g_i$ ([16]).

Cut-HDMR is one method of evaluating the component functions. In cut-HDMR specification of the component functions occurs around anchor points on a regular grid of input points. Define input points around the anchor point as:

$$\begin{aligned} (d_i, \mathbf{D_i^*}) &= (d_1^*, \ldots, d_{i-1}^*, d_i, d_{i+1}^*, \ldots, d_k^*) \\ (d_i, d_j, \mathbf{D_{ij}^*}) &= (d_1^*, \ldots, d_{i-1}^*, d_i, d_{i+1}^*, \ldots, d_{j-1}^*, d_j, d_{j+1}^*, \ldots, d_k^*) \end{aligned} \tag{2.3}$$

The active variables (in red) for the component function vary while all other variables are held at their anchor point value. The functions $g_i$ are then constructed in a recursive fashion.

$$f_0 = f(\mathbf{D}^*, \mathbf{C}^*) \tag{2.4}$$
$$g_i(d_i) = f(d_i, \mathbf{D_i^*}) - f_0 \tag{2.5}$$
$$g_{ij}(d_i, d_j) = f(d_i, d_j, \mathbf{D_{ij}}) - g_i - g_j - f_0 \tag{2.6}$$

The last component function $g_{12\ldots k}$ is found by subtracting off all of the lower dimensional component functions. Finally, a component function is identically zero whenever any input variable is equal to its anchor point value, that is, $d_i = d_i^*$ [16]. The component functions can be defined by a table of values based on a regular grid of the original function evaluations. Interpolation is used to evaluate function values not on the input grid. Each component function represents a smaller dimension, a line, plane, hyperplane, et cetera, cut through the original function.

### 2.3.2 Estimation of $g_D$

By employing some steps from cut-HDMR to estimate $g_D$, the function of discrete inputs, I can exploit the nature of a completely discrete set of inputs since discrete inputs align with how the lower dimensional functions are defined along cut lines, planes, etc. Cut-HDMR techniques have other advantages as well. Enumeration of all discrete input combinations provides a regular grid of inputs. After specification of the anchor point, considerable function evaluations can be avoided, because any component function with $d_i = d_i^*$ is zero. No interpolation is required to evaluate any other points.

Estimation of $g_D$ is summarized in four steps:

1. Choose an anchor point $D^*, C^*$ .
2. Enumerate all possible combinations of the discrete inputs $D$. Eliminate any that contain $d_i = d_i^*$ as an input for a component function. Call this set $\mathcal{D}$.
3. Create a set of samples $(\mathcal{D}, C^*)$ and evaluate the original function at these sample points.
4. Estimate all possible component functions at all possible points.

### 2.3.3 Model choice process

The model choice process proceeds in three steps.

1. Estimate $g_D$ as outlined above.

2. Create the output epdf of $g_D$:

   (a) Sample over the discrete inputs. This should be a large sample and can include any distributional information about the discrete inputs. Epistemic methods could also be applied.

   (b) Evaluate $g_D$ at these inputs.

   (c) Estimate the distribution of the outputs, the epdf of $g_D$.

3. Perform UQ using a method designed for continuous inputs for the transformed function.

   (a) The transformed function is

   $$f(\mathbf{D}, \mathbf{C}) \approx X + g_C \tag{2.7}$$

   (b) $X$ is a random variable distributed as the epdf of $g_D$.

   (c) $g_C = f(\mathbf{D}^*, \mathbf{C})$

   (d) The continuous inputs $C$ retain their original meaning and distributions.

   (e) The last term of equation (2.7) takes its form from the idea of cut-HDMR; the discrete variables are held at their anchor point values while the continuous variables vary.

In essence, $g_D$ captures the discrete variation around the anchor point while $g_C$ captures the continuous variation around the anchor point.

## 2.4 Summary of the conceptual model

We began with an extremely general conceptual model (2.1 and showed one way to apply it to one class of problems generally based on cut-HDMR; however, any estimation procedure or surrogate process can be used to estimate $g_D$ and $g_C$. For example, one could create a multivariate adaptive regression spline (MARS [7]) fit for $g_D$ and a Gaussian process model for $g_D$. There are many places to customize the conceptual model or the process to the problem at hand. A number of suggestions follow:

1. include an interaction term for a subset of $D, C$ when strong interactions are known or suspected
2. use another method to estimate the HDMR component functions, such as RS-HDMR [15] when more of the input space needs to be explored
3. use a lower order set of component functions to estimate $g_D$ (for example, only functions of one or two variables) when the function evaluation budget is too small to evaluate the original function on a full factorial design
4. use a full or lower order HDMR estimate for $g_C$ when the function evaluation budget is small

5. use a constant value (such as the mean) for unimportant variables.

This page intentionally left blank.

# Chapter 3

# Model Choice Example

The function for this example comes from [23], one of a number of mixed variable problems for use in optimization testing. This problem has four discrete binary variables and three continuous variables.

$$\begin{aligned} f(\mathbf{x},\mathbf{y}) &= (y_1-1)^2+(y_2-1)^2+(y_3-1)^2-\ln(y_4+1) \\ &\quad +(x_1-1)^2+(x_2-2)^2+(x_3-3)^2 \end{aligned} \tag{3.1}$$

Subject to: $\mathbf{y} \in \{0,1\}^4 \quad 0 \leq \mathbf{x} \leq 5$ Note that the range for the discrete part of the problem is $[-\ln 2, 3.0]$ and for the continuous part $[0, 29]$. We assign various distributions to the inputs.

## 3.1 Proof of concept example

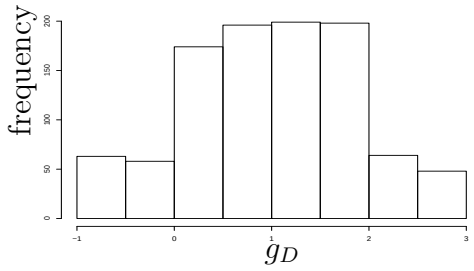
The proof of concept example was used for initial testing of the conceptual model in order to assess the feasibility of transforming a set of discrete inputs into a continuous variable for use in UQ methods designed for continuous inputs. To that end $g_D$ and $g_C$ are not estimated, but are directly assigned to be

$$\begin{aligned} g_D &= (y_1-1)^2+(y_2-1)^2+(y_3-1)^2-\ln(y_4+1) \\ g_C &= (x_1-1)^2+(x_2-2)^2+(x_3-3)^2 \end{aligned}$$

and $f(D,C) = g_D + g_C$ is an exact representation of the original function. For this example, $y_i$ are identically distributed binary random variables with $\Pr(0) = \Pr(1) = 0.5$ and $x_i$ are identically distributed uniform random variables on $[0, 5]$.

We now proceed through steps two and three of the model choice process. For step two, a large set of discrete samples are generated (1000 LHS samples), $g_D$ is evaluated, and the output values are summarized with a histogram. This histogram is the edpf of $g_D$ shown in Figure 3.1. For step 3 the epdf of $g_D$ and the three uniformly distributed $x_i$ are the continuous random inputs to a variety of UQ methods designed for continuous inputs to assess the uncertainty in $X + g_C$.

**Figure 3.1.** Proof of concept example: epdf based on 1000 LHS inputs to $g_D$




Results from the UQ methods designed for continuous inputs on $X + g_D$ are shown in Figure 3.2. The numbers in the key indicate the number of function evaluations. For the new techniques, the function evaluations are written as $1000 + x$, to reflect the 1000 evaluations that were used to estimate the pdf of $g_D$ plus the additional evaluations used by the continuous UQ method. The benchmark cdf of the original function outputs (in pink) was created from exhaustively sampling the original function. The output cdf is estimated as a histogram from 10,000 LHS samples; there really is no other established method that can handle arbitrary distributions on the discrete variables. The other three curves are the result of UQ methods designed for continuous inputs: polynomial chaos expansion (PCE), efficient global reliabilility (EGRA) and local reliability. PCE is a surrogate model using a multidimensional orthogonal polynomial basis and estimated stochastic coefficients [8]. Level 4 quadrature points were used to estimate the coefficients; this adds 1024 function evaluations. The complete cdf (black) shows good agreement with the benchmark LHS. An efficient global reliability method is shown in green (EGRA) [2] which adds 21 function evaluations. EGRA creates a surrogate model using a Gaussian process, refining the model around limit points $x$ in order to apply a reliability method. The reliability method seeks to answer the question, what is the probability the function value exceeds $x$ by transforming the input variables and linearizing about the most probable point. Since this problem is highly nonlinear, the reliability cdf in red gives a notably different result, although it adds the fewest function evaluations. Overall, this example illustrates how the conceptual model can lead to comparable results with fewer functional evaluations.

## 3.2 Example using the full model choice process

This example employs the full model choice process on the problem defined in equation 3.1. Again, $y_i$ are identically distributed binary variables with $\Pr(0) = \Pr(1) = 0.5$ and $x_i$ are identically distributed uniform random variables on $[0, 5]$. The function $g_D$ is created with its full ANOVA decomposition; $g_C$ is defined by the original function with the discrete variables set at their anchor point, i.e., $g_C = f(D^*, C)$. The anchor point is chosen to be

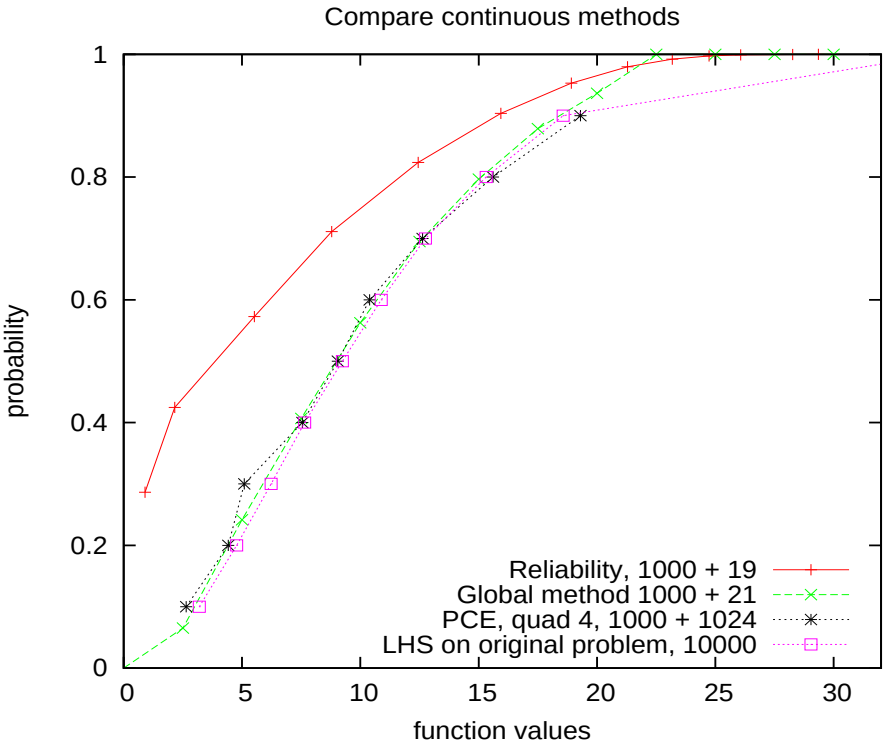



**Figure 3.2.** Proof of concept example

$(D^*, C^*) = (0, 0, 0, 0, 2.5, 2.5, 2.5)$. The three step model choice process is implemented as follows:

### 3.2.1 Estimate $g_D$

The full ANOVA decomposition of $g_D$ is

$$\begin{aligned} g_D(\mathbf{y}) &= g_{D^*} + g_{d1} + g_{d2} + g_{d3} + g_{d4} \\ &\quad + g_{d1d2} + g_{d1d3} + g_{d1d4} + g_{d2d3} + g_{d2d4} + g_{d3d4} \\ &\quad + g_{d1d2d3} + g_{d1d2d4} + g_{d2d3d4} + g_{d1d3d4} + g_{d1d2d3d4} \end{aligned}$$

A component function is identically zero when evaluated at a point such that a component variable takes an anchor point value, for example, $g_{d1d2d3}(0, d2, d3) = 0$. To illustrate the estimation of a single variable component functions, consider $g_{d1}$. Because $d1 = y_1$ is a binary variable, there are only two points to consider. $g_{d1}(0) = 0$ because it is evaluated at an anchor point. $g_{d1}(1) = f(1, 0, 0, 0, 2.5, 2.5, 2.5)$ requiring one function evaluation. This completely determines $g_{d1}$. The other single variable component functions are similarly estimated. These evaluations are used to determine the values of the 16 component functions $\{y_1, y_2, y_3, y_4, y_{12}, y_{13}, \ldots, y_{1234}\}$; the sum of the component functions estimates $g_D$.

### 3.2.2 Estimate the epdf of $g_D$

At this point we have $g_D$, by applying cut-HDMR estimation to a full ANOVA/HDMR decomposition, and we sample it to create the epdf of $g_D$ consistent with the distributions of $\mathbf{D}$. Ths is essentially a mini-UQ that captures the uncertainty in the discrete variables. We advocate a utilizing a large random sample (perhaps ¿100 times the number of discrete inputs) at this step, especially when the discrete choices are equiprobable. The LHS samples can be too uniform, because they are chosen according to an equiprobable bin. The cut-HDMR form of $g_D$ used here is extremely cheap to run, so evaluating a large sample takes very little time.

Large samples are necessary here, because in effect we are averaging over the effect of the discrete variables. To illustrate this point, two epdfs of $g_D$ were created, one with 50 samples and one with 1000 samples. With only 50 samples the lower $g_D$ function values are over-represented and the granularity of the epdf is not well resolved. (show histograms??)

Each of these epdfs are used with the other continuous variables $\mathbf{x}$ as inputs to a PCE estimate of the functional output cdf. Figure 3.4 shows the two estimates of the PCE-created cdf with the benchmark LHS estimate (green). While there is close agreement using the 1000-sample epdf of $g_D$ with PCE (blue cdf); the 50-sample epdf if $g_D$ with PCE shows the underestimates of the discrete effects carries through to the final cdf (red).

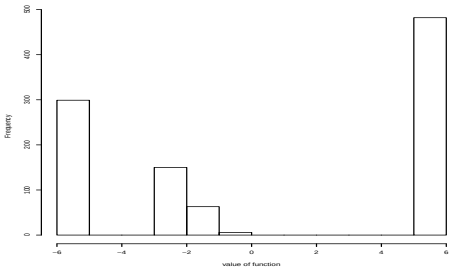



**Figure 3.3.** Epdf of $g_D$ estimated by cut-HDMR

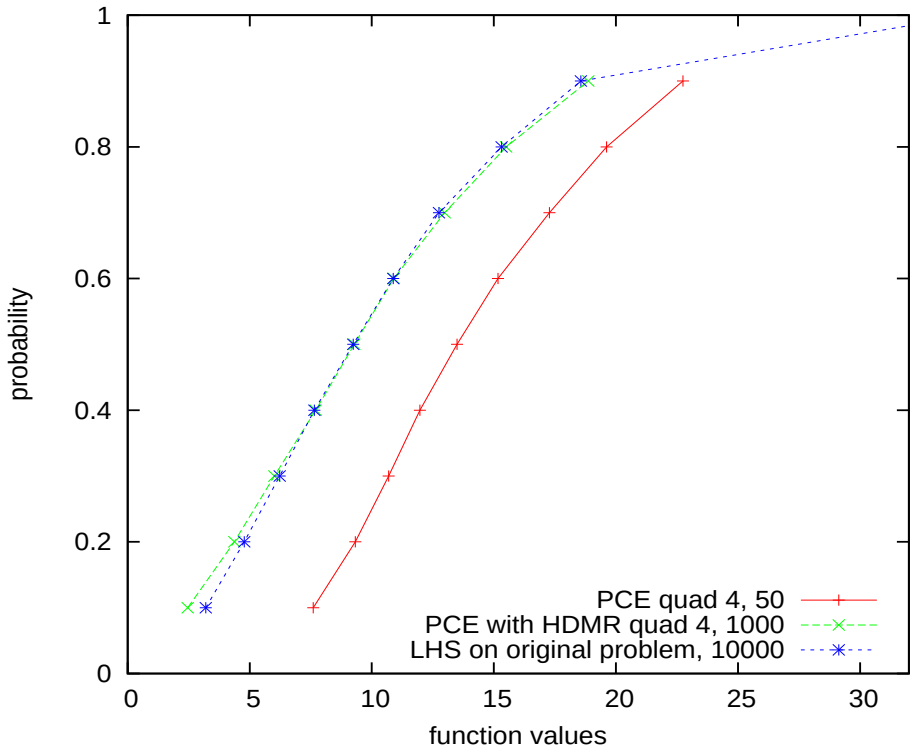



**Figure 3.4.** Insufficient discrete samples: 50 or 1000 samples were used to estimate $g_D$ followed by PCE estimate of the output cdf. The benchmark LHS estimate is included for comparison.

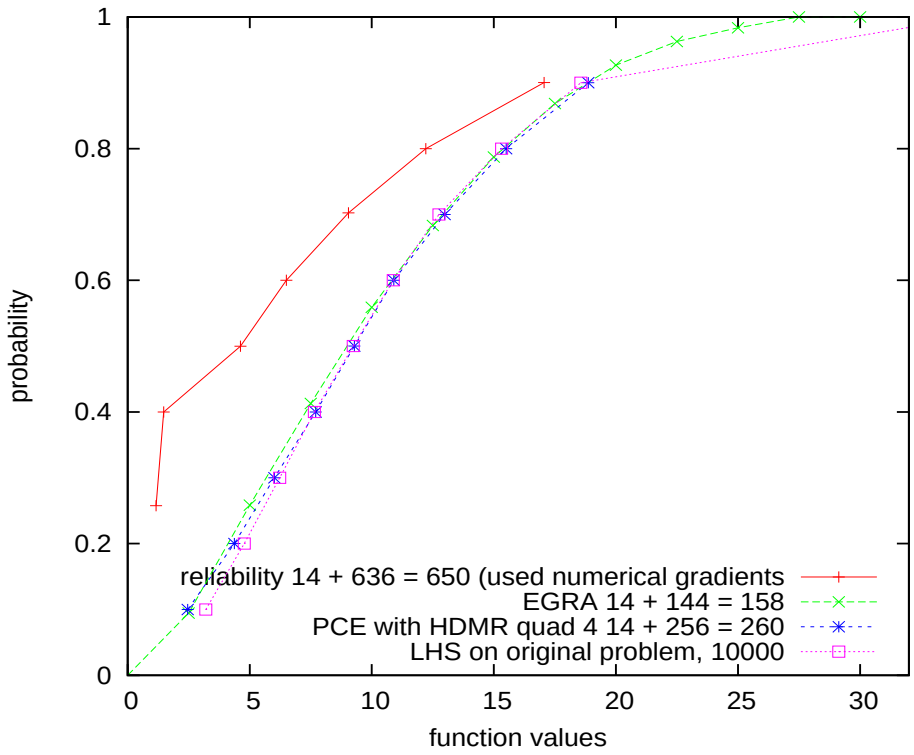



**Figure 3.5.** Compare continuous methods, full model choice process

### 3.2.3 Estimate uncertainty using a UQ method designed for continuous inputs

As in section 3.1 the 1000 sample epdf of $g_D$ and distributions of $\mathbf{x}$ are used in three different UQ methods for continuous inputs and compared to the benchmark LHS result. The results are shown in Figure 3.5 along with the total number of function evaluations (recall that 14 evaluations were used to estimate $g_D$. Against the exhaustive LHS (pink) as the benchmark, PCE (blue) and EGRA (green) show excellent agreement. The reliability method (red) suffers when faced with this highly nonlinear problem; it doesn't even converge for probability levels 0.2 or 0.3.

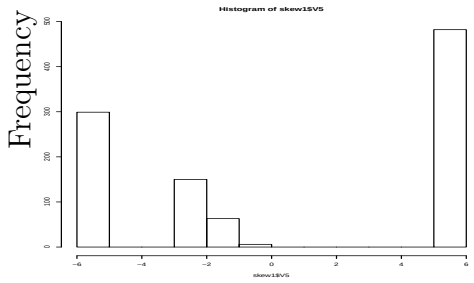



**Figure 3.6.** Epdf of $g_D$ with $\Pr(0) = 1/6$

## 3.3 Example using the full model choice process with different input distributions

Up to this point, all the variables have been uniformly distributed. This example illustrates how this method captures the effect of different distributions on $\mathbf{y}$ and $\mathbf{x}$; $\mathbf{y}$ now has $\Pr(0) = 1/6$ and $\Pr(1) = 5/6$, while $\mathbf{x}$ are normally distributed with means 2.5 and standard deviations $\sigma_1 = 1.0, \sigma_2 = 2.0, \sigma_3 = 3.0$. First, the benchmark LHS analysis is repeated with the new distributions for comparison. The cut-HDMR full ANOVA representation of $g_D$ does not change; nothing in its definition requires knowledge of the input variable distributions. However, estimation of the epdf must be re-done in order to reflect the new distributions of $\mathbf{y}$. 1000 random samples of the discrete variables drawn with respect to their new distributions lead to the epdf of $g_D$ shown in Figure 3.6 which is quite different from the previous histogram in Figure 3.3.

Now the epdf of $g_D$ and the normally distributed $\mathbf{x}$ are used as the continuous inputs to the three UQ methods and the resulting cdfs are shown in Figure 3.7. Once again, the PCE cdf (green) and EGRA (pink) give comparable results to the benchmark LHS cdf (red), while the reliability cdf (blue) fails.

## 3.4 Example with interactions

In this section, we add an interaction term $\alpha$ to the test problem in equation 3.1.

$$\begin{aligned} f(\mathbf{x},\mathbf{y}) &= (y_1-1)^2 + (y_2-1)^2 + (y_3-1)^2 - \ln(y_4+1) \\ &\quad +\alpha(y_4+1)x_3 + (x_1-1)^2 + (x_2-2)^2 + (x_3-3)^2 \end{aligned}$$

where $y_i$ are binary variables with $\Pr(0) = \Pr(1) = 1/2$ and $x_i$ are normally distributed variables with mean 2.5 and standard deviations $1.0, 2.0, 3.0$. The model choice process is

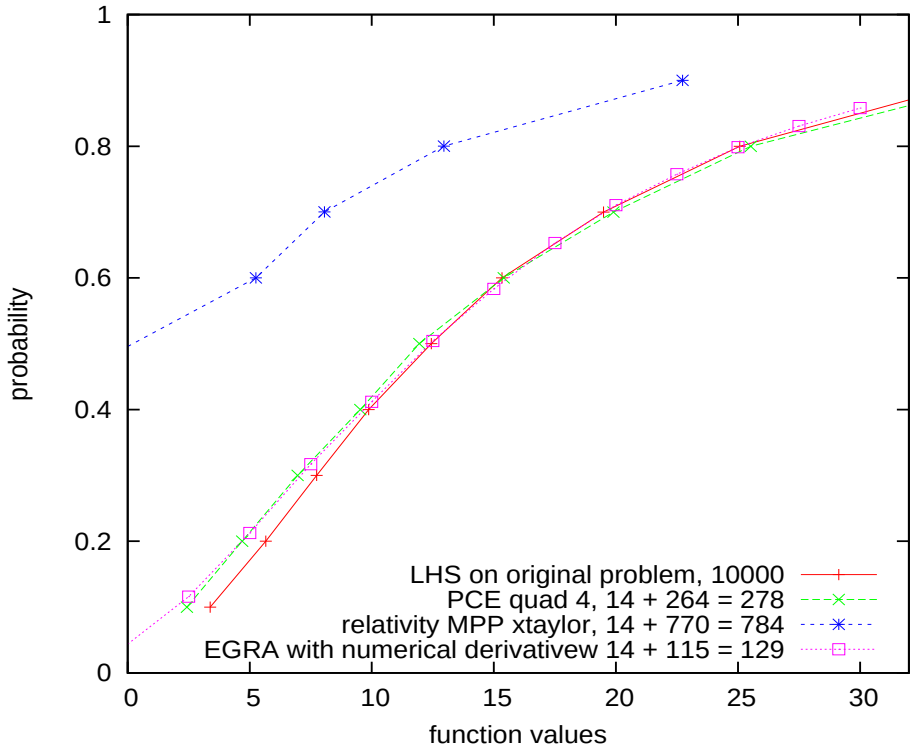



**Figure 3.7.** Compare results for different input distributions

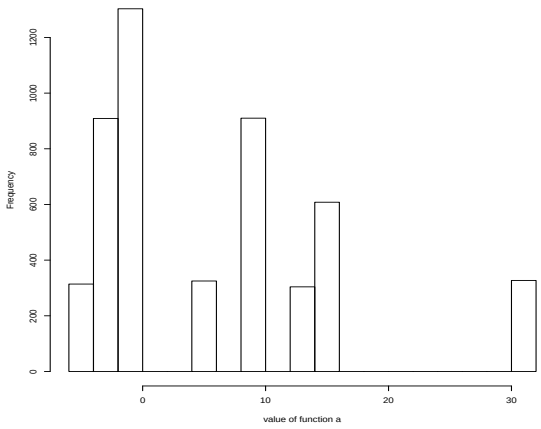



**Figure 3.8.** Epdf of $g_D$ for example with interactions, $\alpha = 1.0$

followed as in previous sections, leading to the estimate of the pdf of $g_D$ shown in Figure 3.8. The effect of the interaction term increases the value of the function 3.2; it also increases the value of $g_D$. Some of the interaction will be detected by this method, but only the interaction between variables and their anchor point values.

Once again, the benchmark 10000 LHS cdf (red) of the original problem with interactions is shown for comparison with the same three UQ methods in Figure 3.7. EGRA (pink) and PCE (green) methods show this method has inflated function values, while the reliability method (blue) underestimates the function values, especially compared to PCE and EGRA. Two possible reasons for these different results are one, that $g_D$ overestimates the effect of the discrete variables on the function (data not shown), and two the variability in $x_3$, one of the interaction terms, is greater that the variability of any of the other inputs. More work is definitely needed to understand functions with interactions.

## 3.5 More on defining model choice problems

During testing, it became clear that certain mixed variable problems are not appropriate for this method.

- periodic problems
- problems where the discrete variable controls the amplitude or reversals in growth such as $y = ax^K$ where $\alpha$ controls the amplitude.

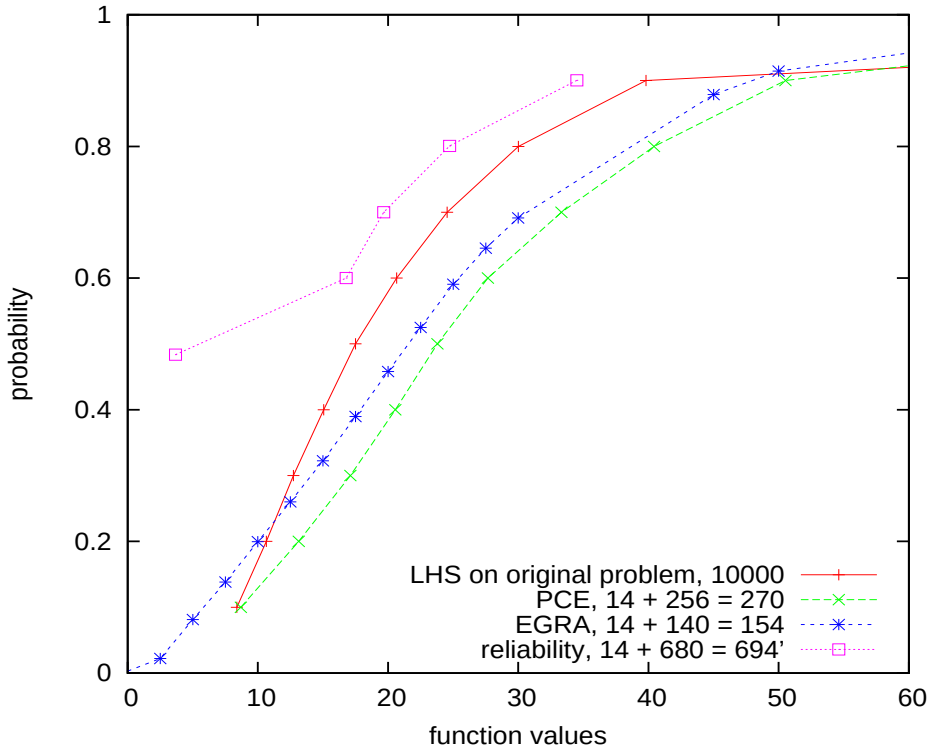



**Figure 3.9.** Compare methods for function with interactions, $\alpha = 1.0$

- problems where the discrete variables are not separable from the continuous variables. One example of this is a pump-and-treat well problem; the discrete variables reflect whether a pump is on or off and the continuous variables describe the location of the well.
- Very little variability due to discrete variables

## 3.6 Convergence

I begin a discussion of convergence with the definition of uniform convergence:

**Definition** We say that a sequences of functions converges *uniformly* on $E$ to a function $f(D, C)$ if for every $\epsilon > 0$ there is an integer $N$ such that $n \geq N$ implies

$$|f_n(D, C) - f(D, C)| \leq \epsilon$$

for all $x \in E$.

[20]

Convergence can be split into two parts: one, convergence of the decomposition 2.1 to the function $f(D, C)$ and two, the convergence of the estimates of the decomposition $g_{D_n}$ and $g_{C_n}$ to $g_D$ and $g_C$. I know that for functions with interactions between $D$ and $C$, the estimates $f_n(D, C)$ fail to uniformly converge to $f(D, C)$ since there is no mechanism to capture the effect of the interactions. However, it is still possible for the estimates of $g_{D_n}$ and $g_{C_n}$ to converge uniformly to $g_D$ and $g_C$; this will be based on the convergence properties of the algorithm used to estimate $g_{D_n}$ and $g_{C_n}$. For functions without such interactions, uniform convergence is achieved for both the function $f_n(D, C)$ and the parts $g_{D_n}$, $g_{C_n}$ based on the convergence properties of the estimation procedure.

How much is lost when there are interactions between the discrete and continuous inputs? Sobol' provides some guidance in [25], since forcing the additive decomposition in 2.1 can be thought of as the factor fixing approach in sensitivity analysis where either the discrete or the continuous inputs are fixed. For a set of inputs fixed at $z_0$, the approximation error depends on $z_0$ :

$$\delta(z_0) = \frac{1}{\mathrm{Var}(f(D, C))} \int [f(D, C) - f(D, z_0)]^2 dD$$

assuming $f(D, C)$ is square integrable and the expected error is of the same order as the total sensitivity $S_z^{tot}$ for the group of inputs $z_0$, i.e., $\mathrm{E}\left(\delta(z_0)\right) = 2S_z^{tot}$.

talk about convergence of cutHDMR?? and convergence of HDMR? needed here??

the number of samples needed for representation to a given tolerance is invariant to the dimensionality of the function, thereby providing for a very efficient means to perform high dimensional interpolation.

[21].

# Chapter 4

# Discussion

At the heart of creating surrogate models or response surfaces is continuity. The true values of the original function can be ignored in some areas, because they can be approximated by a surface informed by a few functional inputs and outputs. The original function does not vary too far from the approximation made through these points; if it does, we can evaluate more input points near the evaluated points. Even if the output is discontinuous or discrete [18, 9], we can exploit the continuity of the inputs to find and assess these jumps. Not so with the discrete inputs; in fact, traditionally in statistics, the discrete covariates fundamentally change the response—in regression-type problems: the mean or the slope, in classification-type problems: the classification [3]. With complicated models, the situation is not so straightforward. This brings difficulties to both types of UQ methods: better sampling strategies will omit some discrete combinations and the creation of surrogates or response surfaces generally fails with discete inputs (or the software does not accept them). Some work is beginning to address the lack of surrogates for mixed discrete-continuous problems [9, 29, 19, 27], including a bioinformatics method for protein engineering that employs some discrete input HDMR modeling [6]. Reliability methods for mixed discrete-continuous problems [11] are also being developed as the need to do UQ on these problems becomes more common. I have presented one method for estimating the functions in the additive decomposition; many others can be used, especially in order to taylor the estimation procedure to the function evaluation budget and known input variable interactions.

I have shown that the idea of transforming a set of discrete variables into a continuous probabilistic one is a powerful method for UQ in model choice problems. Another class of problems where this method may be pertinent are repeated units problems. These problems are often critical infrastructure or logistics problems that occur in a high dimensional space with many discrete and continuous variables; in fact the discrete variables may outnumber the continuous ones. The nature of the probabilistice transform in this class would be to transform the counts into a related continuous quantity. For example, the number of nodes in an electric grid would be transformed into the amount of wire needed for that number of nodes; on or off variables could be transformed into total rate produced; the number of trucks in a platoon could be replaced by the cost to place and use a truck. In cases where one seeks to optimize a cost function, replacing numbers of items with costs would be especially effective.

This type of transformation opens a path to a plethora of possibilities for future work such as:

1. Employing different procedures for estimating $g_D$ and $g_C$ including other ways to use HDMR
2. Using less than full factorial discrete inputs to estimate $g_D$
3. Testing this method on model choice problems with interactions and on a real world problem
4. Use sparse grids and stochastic collocation to estimate $g_D$ and $g_C$. This would dovetail nicely with the discretization of the sparse grid.
5. Investigate the use of this kind of decomposition or transformation in sensitivity analysis
6. Analyze some repeated units problems
7. Investigating other classes or problems and which probabilistic transformations would be appropriate

Behind all the details of distributions and decompositions lies the novel concept of this report: transformation of discrete variables into associated continuous variables for use in UQ methods designed for continuous inputs. This was motivated by problems in statistical genetics where the effect of *discrete* genotypes on the trait of interest are represented as additive *continuous* random variables [1]. The concept opens a wide playing field for transformations and application and combination of different surrogates and sampling techniques. For example, using the

In summary, a collection of discrete variables is transformed into a continuous, probabilistic variable by estimating the joint effect of the variables on the output function for use in uncertainty quantification. UQ can now be carried out one time rather than multiple times for each combination of discrete variables. All uncertain variables are included without assuming isotropic behavior across all unique discrete combinations. The discrete variables are not treated as if they were continuous ones, but naturally transformed into a probabilistic continuous variable. While some iteration over discrete combinations is still required, the potential savings in function evaluations provided by efficient UQ methods designed for continuous inputs can be tremendous.

# Appendix

Let

$$f^+ = \max(f, 0) \quad f^- = -\min(f, 0)$$

**Definition** Let $f$ be measurable and consider the two integrals

$$\int_E f^+ d\mu \quad \int_E f^- d\mu$$

If both integrals are finite, we say that $f$ is *integrable* on $E$ in the Lebesque sense with respect to $\mu$.

A function is *square integrable* if $f^2$ is integrable.

This page intentionally left blank.

# Bibliography


[1] Lara E Bauman, Laura Almasy, John Blangero, Ravi Duggirala, Janet S Sinsheimer, and Ken Lange. Fishing for pleiotropic QTLs in a polygenic sea. *Ann of Hum Gen*, 69(5):590–611, 2005.

[2] Barron J Bichon, Michael S Eldred, Laura Painton Swiler, Sandaran Mahadevan, and John M McFarland. Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA Journal*, 46(10):24592468, 2008.

[3] Leo Breiman, Jerome Friedman, Charles J Stone, and R A Olshen. *Classification and Regression Trees*. Chapman and Hall, 1984.

[4] Richard Li-Yang Chen, Duncan Callaway, and Amy Cohn. Including wind in power system siting and capacity expansion models. manuscript in preparation, 2007.

[5] Rajib Chowdhury and Sondipon Adhikari. Fuzzy parametric uncertainty analysis of linear dynamical systems: A surrogate modeling approach. *Mech Sys Signal Processing*, 32:5–17, 2012.

[6] Xiaojiang Feng, Joaquin Sanchis, Manfred T Reetz, and Herschel Rabitz. Enhancing the efficiency of directed evolution in focused enzyme libraries by the adaptive substituent reordering algorithm. *Chem Eur J*, 2012. advance online release.

[7] Jerome H Friedman. Multivariate adaptive regression splines. *Ann Statist*, 19(1):1–67, 1991.

[8] Roger Ghanem and John Red-Horse. Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element approach. *Physica D*, 133:137144, 1999.

[9] Robert B Gramacy and Herbert K H Lee. Bayesian treed gaussian process models with an application to computer modeling. *J Am Stat Assoc*, 103(483), 2008.

[10] Michael Griebel. Sparse grids and related approximation schemes for higher dimensional problems. In *Foundations of computational mathematics, London Math. Soc. Lecture Note Ser.*, volume 331. Cambridge Univ. Press, Cambridge, 2005.

[11] Subroto Gunawan and Panos Y Papalambros. Reliability optimization with mixed continuous-discrete random variables and parameters. *Trans ASME*, 129:158–165, 2007.

[12] J C Helton and F J Davis. Sampling-based methods for uncertainty and sensitivity analysis. Technical Report SAND99-2240, Sandia National Laboratories, Albuquerque, NM, 2000.

[13] J C Helton, J C Johnson, C J Sallaberry, and C B Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. Technical Report SAND2006-2901, Sandia National Laboratories, June 2006.

[14] Berfin Kalay and Metin Demiralp. Fundamental elements of vector enhanced multivariance product representation. In *AIP Conf Proc*, volume 1479. American Institute of Physics, 2012.

[15] Genyuan Li, Carey Rosenthal, and Herschel Rabitz. High dimensional model representations. *J Phy Chem*, 105(33):7765–7777, 2001.

[16] Genyuan Li, Sheng-Wei Wang, and Herschel Rabitz. High dimensional model representations HDMR: concepts and applications. Technical report, Princeton University, .

[17] Genyuan Li, Sheng-Wei Wang, Carey Rosenthal, and Herschel Rabitz. High dimensional model representations generated from low dimensional data samples. I. mp-Cut-HDMR. *J Math Chem*, 30(1), 2001.

[18] Martin Meckesheimer, Russell R Barton, Timothy Simpson, Frej Limayem, and Bernard Yannou. Metamodeling of combined discrete/continuous responses. *AIAA Journal*, pages 1950–1959, 2001.

[19] Peter Z G Qian, Huaiqing Wu, and C F Jeff Wu. Gaussian process models for computer experiments with qualitative and quantitative factors. author communication, 2007.

[20] Walter Rudin. *Principles of mathematical analysis*. McGraw-Hill, Inc., 1953.

[21] Ömer F Aliş and Herschel Rabitz. Efficient implementation of high dimensional model representations. *J Math Chem*, 29(2), 2001.

[22] Andrea Saltelli, Karen Chan, and E Marian Scott, editors. *Sensitivity Analysis*. John Wiley & Sons, ltd, 2000.

[23] Isaac Siwal. A note on mixed variable mathematical programs, 1997.

[24] Ilya M Sobol'. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math and Computers in Simulation*, 55:271–280, 2001.

[25] Ilya Meerovich Sobol', Stefano Tarantola, Sergei S Kucherenko, and Wolfgang Mauntz. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Rel Eng & Sys Safety*, 92:957–960, 2007.

[26] R Srinivasan. *Importance Sampling*. Springer-Verlag, 2002.

[27] Curtis B Storlie, Howard D Bondell, Brian J Reich, and Hao Helen Zhang. Surface estimation, variable selection, and the nonparameteric oracle property. from LANL, 2009.

[28] Laura P Swiler and A A Giunta. Aleatory and epistemic uncertainty quantification for engineering applications. Technical Report SAND2007-2670C, Sandia National Laboratories, 2007.

[29] Qiang Zhou, Peter Z G Qian, and Shiyu Zhou. A simple approach to emulation for computer models with qualitative and quantitative factors. in preparation, 2010.

This page intentionally left blank.

Sandia National Laboratories