Experimental Design of Formulations Utilizing High Dimensional Model Representation

DATASET · AUGUST 2015

READS
13

4 AUTHORS, INCLUDING:

Herschel Rabitz
Princeton University
947 PUBLICATIONS **23,825** CITATIONS

SEE PROFILE

THE JOURNAL OF PHYSICAL CHEMISTRY A

Article

pubs.acs.org/JPCA

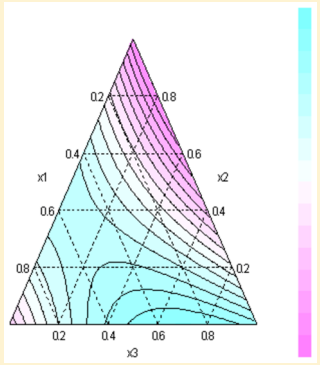
Experimental Design of Formulations Utilizing High Dimensional Model Representation

Genyuan Li,[†] Caleb Bastian,[†] William Welsh,[‡] and Herschel Rabitz*,[†]

[†]Department of Chemistry, Princeton University, Princeton, New Jersey 08544, United States

[‡]Department of Pharmacology, Robert Wood Johnson Medical School, and Division of Cheminformatics, Biomedical Informatics Shared Resource, Cancer Institute of New Jersey at Rutgers University, Piscataway, New Jersey 08854, United States

ABSTRACT: Many applications involve formulations or mixtures where large numbers of components are possible to choose from, but a final composition with only a few components is sought. Finding suitable binary or ternary mixtures from all the permissible components often relies on simplex-lattice sampling in traditional design of experiments (DoE), which requires performing a large number of experiments even for just tens of permissible components. The effect rises very rapidly with increasing numbers of components and can readily become impractical. This paper proposes constructing a single model for a mixture containing *all* permissible components from just a modest number of experiments. Yet the model is capable of satisfactorily predicting the performance for full as well as all possible binary and ternary component mixtures. To achieve this goal, we utilize biased random sampling combined with high dimensional model representation (HDMR) to replace DoE simplex-lattice design. Compared with DoE, the required number of experiments is significantly reduced, especially when the number of permissible components is large. This study is illustrated with a solubility model for solvent mixture screening.



1. INTRODUCTION

Formulations or mixtures of multiple components arise in many applications from catalysts to alloys and solvents. This paper will present a new procedure for handling such problems, especially when large numbers of components can arise. As a specific example, we will illustrate the tools for solvent mixture screening, which presents a significant challenge when seeking to discover optimal solvent mixtures with only a few components out of a large possible set.[1−3]

Although binary and ternary mixtures are often used in practice, the number of permissible components may be as large as tens or even hundreds. Finding suitable binary or ternary mixtures with desired properties from all possible binary and ternary combinations among the permissible components by traditional design of experiments (DoE) requires a large number of experiments. The problem is just as complex, or even more so, when searching for an optimal mixture of more than three components where the number of possible combinations could be enormous.

Suppose $x_i \geq 0$ $(i = 1, 2, ..., n)$ denotes the fraction of the ith component, then for a mixture with n components, we have

$$\sum_{i=1}^{n} x_i = 1 \tag{1}$$

All such points, $\mathbf{x}$, in the n-dimensional space compose an $(n - 1)$-dimensional simplex. In DoE, simplex-lattice designs are often used for selecting mixtures with a maximum of m components drawn from n possible components. An $\{n, m\}$ simplex-lattice design consists of points defined by the following coordinate settings: the fractions assumed by each component take on $m + 1$ equally spaced values from 0 to 1,

$$x_i = 0,\ 1/m,\ 2/m,\ ...,\ 1, \quad \text{for } i = 1, 2, ..., n$$

and all possible combinations (mixture formulations) are considered for assessment. In this setting a mixture can only contain less than or equal to m components. The number of design points in the simplex-lattice is[4]

$$(n + m - 1)!/(m!(n - 1)!) \tag{2}$$

Consider a three-component mixture for which the number of equally spaced levels for each component is four (i.e., $x_i = 0, 1/3, 2/3, 1$). In this example $n = 3$ and $m = 3$. If one uses all possible blends of the three components with these fractions, the $\{3, 3\}$ simplex-lattice design then contains 10 blending coordinates or samples (see the dots in Figure 1). If there are $n = 10$, 20, or 50 permissible components, to find the best desired binary or ternary mixtures, we have to test all possible two and three component combinations. From eq 2, we see that the required numbers of experiments by the $\{n, 3\}$ simplex-lattice design are 220, 1540, and 22 100, respectively. As n gets larger, performing such a growing number of experiments becomes increasingly prohibitive.

DoE usually constructs a specific model as a simple polynomial for each m component mixture. For example,

Received: May 22, 2015
Revised: June 17, 2015
Published: June 19, 2015

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

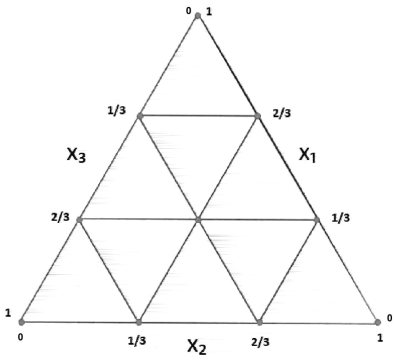


Figure 1. {3, 3} simplex-lattice design often utilized in standard DoE.

with a {3, 3} simplex-lattice design, the following model may be used

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{1 \le i < j \le 3} \beta_{ij} x_i x_j + \beta_{123} x_1 x_2 x_3 \quad (3)$$

where y denotes the relevant property of the mixture (e.g., y = ln S, the logarithmic value of solute solubility S in a solvent mixture), and the unknown parameters β_0, β_i, β_{ij}, and β_{123} are determined by regression from the 10 experimental observations. The resultant model is then used to search for the optimal composition among the three components.

To overcome the shortcoming of the DoE $\{n, m\}$ simplex-lattice design with typically $n \gg m$, we propose a new method: instead of directly constructing every m component mixture model, we first construct a single model for a mixture with *all* n permissible components present, while just using a modest number of experiments. The feasibility of this method resides on the following: (1) the construction of a mathematical model involving tens of variables is generally feasible with modern methodology; (2) if the model correctly reflects the desired property of the full formulation, it should satisfactorily predict the property for any composition with full or a reduced number of components including the cases with as few as one, two or three components. In order to be successful, the new experimental design must only require a modest number of experiments while still producing an accurate model with n variables, where n may be tens or more. The proposed method presented in this paper utilizes (a) a special form of biased random sampling to replace the simplex-lattice design along with (b) high dimensional model representation (HDMR) to replace the simple polynomial models of standard DoE. We will show that this method permits a significantly reduced sample size, while still producing satisfactory predictions for full component mixtures as well as all possible binary and ternary mixtures. This demonstration will be performed with a solvent mixture model, relevant to crystallization efforts (e.g., in the purification of pharmaceuticals).

The paper is organized as follows. In section 2 various sampling strategies will be compared. Section 3 briefly summarizes the principles of HDMR. The details of the HDMR methodology are given in the Appendices. An example with 10 solvents is used in section 4 for illustration. Finally, section 5 contains concluding remarks.

2. SAMPLING STRATEGIES

Without any prior knowledge of the composition property $f(\mathbf{x})$ of the mixture, where $\mathbf{x} = (x_1, x_2, ..., x_n)$ denotes the component fractions, then uniform sampling of $\mathbf{x}$ in the simplex might appear to be the best choice to determine $f(\mathbf{x})$. It can be proved that uniform sampling in a simplex corresponds to sampling from a Dirichlet distribution with all the Dirichlet parameters α_i having the value 1.[5−7]

The Dirichlet distribution of order $n \ge 2$ (i.e., a mixture with $n \ge 2$ components) with parameters $\alpha_1, ..., \alpha_n > 0$ has a probability density function with respect to a Lebesgue measure on the Euclidean space $\mathbb{R}^{n-1}$ given by[6]

$$f(x_1, ..., x_{n-1}; \alpha_1, ..., \alpha_n) = \frac{1}{B(\alpha)} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \quad (4)$$

over the open $(n - 1)$-dimensional simplex defined by

$$x_1, \cdots, x_{n-1} > 0,$$
$$x_1 + ... + x_{n-1} < 1,$$
$$x_n = 1 - x_1 - ... - x_{n-1}$$

and zero elsewhere. The normalizing constant $B(\alpha)$ is the multinomial beta function, which can be expressed in terms of the gamma function:

$$B(\alpha) = \prod_{i=1}^{n} \Gamma(\alpha_i) / \Gamma\left(\sum_{i=1}^{n} \alpha_i\right), \quad \alpha = (\alpha_1, ..., \alpha_n) \quad (5)$$

Figure 2 gives the projection of 1000 random samples into the (x_1, x_2)-subspace obtained by uniform sampling in the simplex with n = 10 variables.

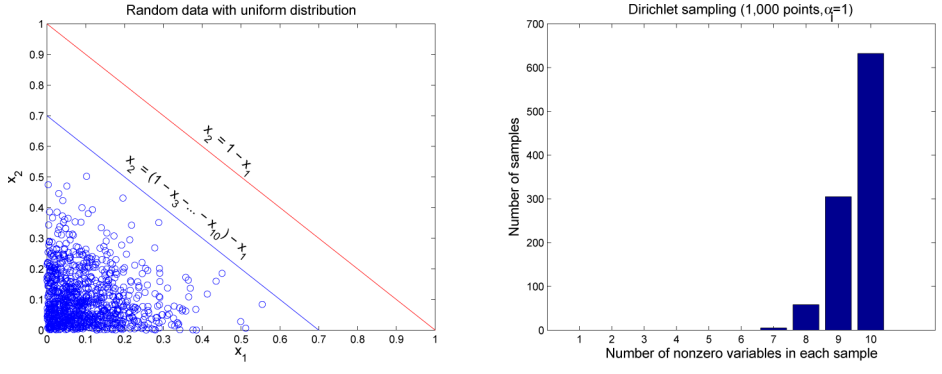


Figure 2. Uniform sampling (i.e., Dirichlet sampling with all α_i = 1) distribution with n = 10. Left panel: projection into the (x_1, x_2)-subspace. Right panel: distribution for the number of nonzero variables.

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

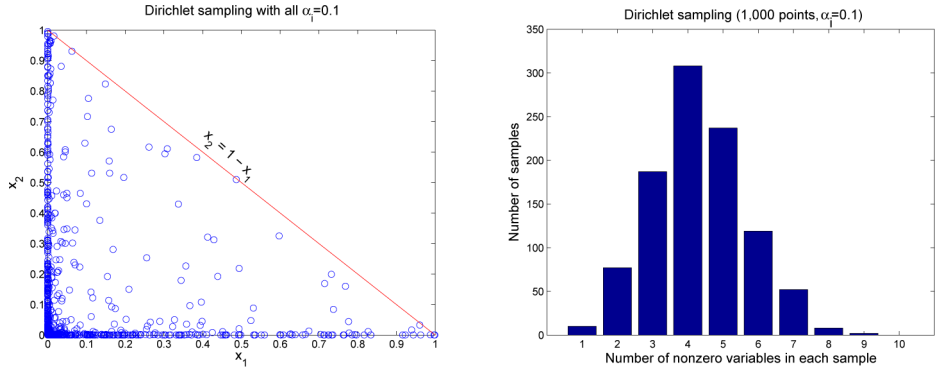


Figure 3. Dirichlet sampling with all α_i = 0.1. Left panel: projection into the (x_1, x_2)-subspace. Right panel: distribution for the number of nonzero variables.

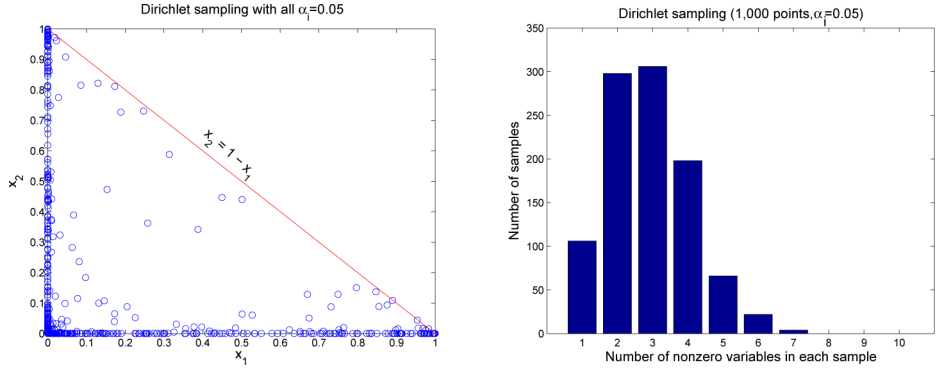


Figure 4. Dirichlet sampling with all α_i = 0.05. Left panel: projection into the (x_1, x_2)-subspace. Right panel: distribution for the number of nonzero variables.

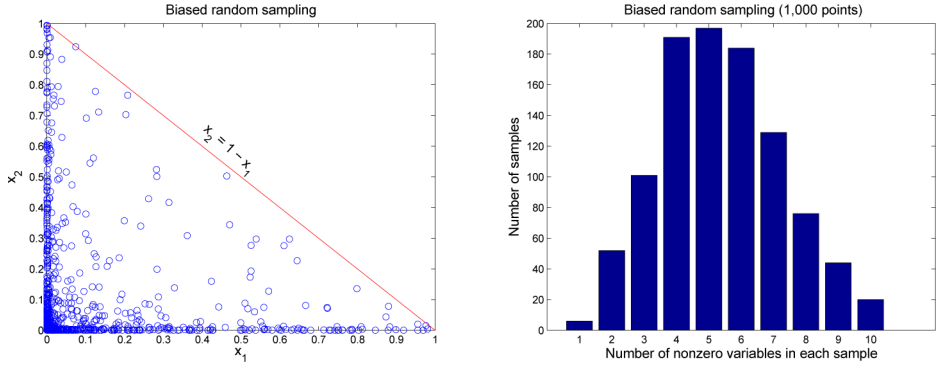


Figure 5. Distribution of biased random sampling. Left panel: projection into the (x_1, x_2)-subspace. Right panel: distribution for the number of nonzero variables.

Note that for uniform sampling, the points are pressed to the lower left corner of the (x_1, x_2)-subspace. This is easy to understand because the blue line represents

$$x_2 = b - x_1 = (1 - \sum_{j=3}^{10} x_j) - x_1 \quad (6)$$

having an intercept b which becomes even smaller than 1 with larger n when x_i $(i = 1, 2, ..., n)$ are uniformly sampled. Taking a practical perspective, we will consider two digits of accuracy for the fractional components, and treat a fraction less than 0.005 as zero, then the right panel in Figure 2 shows that for n = 10 and 1000 points, uniform sampling does not provide information on boundary compositions of dimension less than seven components. This behavior implies that uniform sampling does not adequately cover the lower dimensional boundaries (i.e., formulations) of the simplex. Knowledge of these lower dimensional boundaries is often the goal corresponding to seeking a few variable final objective function.

Fortunately, Dirichlet sampling is flexible by changing the parameters α_i. When all the parameters are set at α_i = 0.1, the sampling moves toward boundaries. When all α_i = 0.05, even more data are sampled on or close to the boundaries. Figures 3 and 4 give the Dirichlet sampling with α_i = 0.1, 0.05, respectively.

An alternative procedure is to perform biased random sampling, which may be done in several ways and here we consider the following:

(1) Uniformly sample each of the n variables $\hat{x}_i$ $(i = 1, 2, ..., n)$

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

$$\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_n), \quad \hat{x}_i \in [0, 1]$$

(2) Perform a random permutation $\mathbf{p}$ of 1, 2, ..., n:

$$\mathbf{p} = (p_1, p_2, ..., p_n)$$

3) Construct new random variables x_i (i = 1, 2, ..., n) from $\hat{\mathbf{x}}$:

$$x_{p_1} = \hat{x}_{p_1},$$
$$x_{p_2} = \hat{x}_{p_2}(1 - x_{p_1}),$$
$$\cdots$$
$$x_{p_{n-1}} = \hat{x}_{p_{n-1}}(1 - x_{p_1} - ... - x_{p_{n-2}}),$$
$$x_{p_n} = 1 - x_{p_1} - ... - x_{p_{n-1}}$$

(4) Repeat procedures 1 to 3 to generate N data points of $\mathbf{x}$.

Thus, every variable x_i has an equal chance of taking position 1, 2, ..., n. Figure 5 gives the distribution of biased random sampling with n = 10 and 1000 points. Compared to Dirichlet sampling with α_i = 1, 0.1, 0.05, the distribution of biased random sampling given on the right panel of Figure 5 is closer to a normal distribution with mean $\sim n/2 = 5$. This form of biased random sampling possibly corresponds to Dirichlet sampling with proper values for the α_i's, but the advantage of biased random sampling is that we do not need to search for the α_i parameters.

We seek to test these sampling methods and find the one which gives the best prediction quality for full component mixtures as well as binary and ternary mixtures with the smallest number of samples in the full mixture of n components. This test will be performed in section 4 with a solubility model. Section 3 first summarizes the HDMR technique used to represent the property $f(\mathbf{x})$ over the full space of all mixture components.

3. PRINCIPLES OF HDMR

To construct a single model describing the composition property $f(\mathbf{x})$, a simple global polynomial

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{1\le i<j\le n} \beta_{ij} x_i x_j + \cdots + \sum_{1\le i_1<...<i_l\le n} \beta_{i_1 i_2 ... i_l} \prod_{j=1}^{l} x_{i_j} + \cdots + \beta_{12...n} \prod_{i=1}^{n} x_i \tag{7}$$

is generally not adequate. Furthermore, the total number k of unknown parameters $\{\beta\}$ for the polynomial with n variables is

$$k = \sum_{r=0}^{n} C_n^r = 2^n \tag{8}$$

where C_n^r denotes the number of all combinations of n things taken r at a time. k can be a very large number for a large value of n. To accurately determine these unknown parameters by regression, more than k data points are often needed which implies that a large number of experiments must be performed. Therefore, a simple global polynomial is likely not suitable for constructing a property model with a large value for n.

In contrast to eq 7, HDMR is a general set of quantitative model assessment and analysis tools for capturing high dimensional input-output system behavior,[8−12] and it may be employed to efficiently construct a model with all permissible components. Here, the principles of HDMR are briefly introduced. The details of HDMR methodology are given in the Appendices.

Many problems in science and engineering reduce to the need for efficiently constructing a map of the relationship between a set of high dimensional system inputs $\mathbf{x}$ and the system output $f(\mathbf{x})$. As the contributions of the multiple input variables upon the output can be independent and cooperative, it is natural to express $f(\mathbf{x})$ as a finite hierarchical expansion:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i, x_j) + \cdots + f_{12...n}(x_1, x_2, ..., x_n) = \sum_{u\subseteq n} f_u(\mathbf{x}_u) \tag{9}$$

Here we use the following multi-index notation. Given the subset $u \subseteq \{1, 2, ...,n\}$, we denote by $\mathbf{x}_u$ those variables in $\mathbf{x}$ whose indexes are in u. Note that the empty set $\varnothing$ is a subset of $\{1, 2, ..., n\}$ and we have $f_\varnothing = f_0$, a constant. We will also write $u \subseteq n$ in place of $u \subseteq \{1, 2, ..., n\}$ for simplicity.

The expansion given in eq 9 has been known for sometime, and often referred to as the ANOVA decomposition.[13,14] The key operational issue is how to construct the component functions $f_u(\mathbf{x}_u)$ in eq 9. One advantage of HDMR is that these component functions are *optimally* and *uniquely* constructed to maximize the contribution of low order component functions such that $f(\mathbf{x})$ often can be satisfactorily approximated by truncated low order HDMR expansions,[8−10] e.g., the second order HDMR expansion

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i, x_j) \tag{10}$$

or the third order HDMR expansion

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{1\le i<j\le n} f_{ij}(x_i, x_j) + \sum_{1\le i<j<k\le n} f_{ijk}(x_i, x_j, x_k) \tag{11}$$

The uniqueness of the HDMR component functions is guaranteed by the *mutual orthogonality* condition for independent variables

$$\mathbb{E}[f_u(\mathbf{x}_u) f_v(\mathbf{x}_v)] = 0, \quad v \neq u \tag{12}$$

or *hierarchical orthogonality* condition for correlated variables

$$\mathbb{E}[f_u(\mathbf{x}_u) f_v(\mathbf{x}_v)] = 0, \quad v \subset u \tag{13}$$

Here, $\mathbb{E}[\cdot]$ denotes the expectation value.[11,12] For the mixture problem considered in this work, the variables are correlated due to the relation in eq 1. Other formulation problems (e.g., components added to a large volume of solvent) may have no constraints on the components. Thus, we include here the cases of HDMR with independent or correlated variables.

Analytical forms for HDMR component functions may not be readily obtained for an arbitrary function $f(\mathbf{x})$ along with an arbitrary probability distribution of $\mathbf{x}$. However, in practice, the HDMR component functions may be approximated by some suitable basis functions $\phi_{uk}(\mathbf{x}_u)$

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

$$f_u(\mathbf{x}_u) \approx \sum_{k=1}^{s_u} c_{uk}\phi_{uk}(\mathbf{x}_u), \qquad \varnothing \neq u \subseteq n \tag{14}$$

and

$$f(\mathbf{x}) = f_0 + \sum_{\varnothing \neq u \subseteq n} \sum_{k=1}^{s_u} c_{uk}\phi_{uk}(\mathbf{x}_u) \tag{15}$$

where the c_{uk}'s are constant combination coefficients, s_u is an integer, and the $\phi_{uk}(\mathbf{x}_u)$'s are polynomials, splines, etc.[9]

As the mixture variables are correlated by the relation in eq 1, the basis functions $\phi_{uk}(\mathbf{x}_u)$ and their combination coefficients c_{uk} must be chosen in such a way that the HDMR component functions are *hierarchically orthogonal*.[11] Therefore, ideally, the basis functions used for high order HDMR component functions should be orthogonal to the basis functions used in its nested lower order HDMR component functions. For a second order HDMR expansion, for example, the basis functions used for $f_{ij}(x_i, x_j)$ should be orthogonal to the basis functions used for $f_i(x_i)$ and $f_j(x_j)$. If $\phi_k^{(i)}(x_i)$, $\phi_l^{(j)}(x_j)$ ($k, l = 1, 2, ...$) are basis functions for $f_i(x_i)$ and $f_j(x_j)$, then $\phi_m^{(ij)}(x_i, x_j)$ ($m = 1, 2, ...$) used for $f_{ij}(x_i, x_j)$ must be orthogonal to all of them. Such a relation is the foundation for defining multivariate orthogonal polynomials. A general approach to construct orthogonal polynomials of several variables with an arbitrary probability distribution for the variables has been developed.[15] However, the direct application of multivariate orthogonal polynomials to construct HDMR component functions has two drawbacks: (1) to construct the basis functions used for a high order HDMR component function, the degree of the polynomial basis functions used for its nested lower order HDMR component functions must be known in advance. Improper setting of the degree may cause a large error for the HDMR model; (2) for large n the number of polynomial basis functions will in turn also be large. For example, for the second order HDMR expansion with 10 variables, the total number of polynomial basis functions (consequently, the number of unknown parameters) can be 370. To determine all of these parameters by regression, more than 370 experimental data are needed. Rahman[16] proposed using the multivariate orthonormal polynomials as basis functions for the construction of the HDMR component functions by solving the coupled system of equations satisfying the hierarchical orthogonal condition of the component functions. This method has the same two drawbacks.

A numerical method,[11,12] a combination of extended bases[17] and D-MORPH (diffeomorphic modulation under observable response preserving homotopy) regression,[18,19] has been developed to properly construct the basis functions and accurately determine their combination coefficients c_{uk}. A summary of the method is given below, and the details can be found in the Appendices and refs 11, 18, and 19.

- This method is capable of constructing mutually (for independent variables) or hierarchically (for correlated variables) orthogonal HDMR component functions from various basis functions with respect to the probability distribution of $\mathbf{x}$.
- The basis functions used in a high order HDMR component function contain all the basis functions used in its nested lower order HDMR component functions. Such a structure is referred to as employing *extended basis* functions, which guarantee the hierarchical orthogonality of the HDMR component functions when the variables are correlated.
- When the basis functions are polynomials, instead of constructing multivariate orthogonal polynomials first from low order to high order step-by-step (which requires properly setting the polynomial degree at each step), this method simultaneously constructs the different order HDMR component functions represented as expansions of the chosen basis functions by D-MORPH regression. *The latter procedure forces the* HDMR *component functions to be hierarchically orthogonal with respect to a given probability distribution*. In the case of polynomial basis functions, the resultant HDMR expansion is the multivariate hierarchically orthogonal polynomial expansion. For independent variables, the hierarchically orthogonal polynomial expansion automatically reduces to the mutually orthogonal polynomial expansion.
- This method not only correctly constructs the HDMR component functions with independent and/or correlated variables, but also has an extra advantage that the required data can be *less* than the number of unknown parameters without the precondition that the unknown parameters are sparse. This advantage is especially useful for reducing the number of experiments.

4. ILLUSTRATION: APPLICATION TO A SOLVENT MIXTURE MODEL

The Jouyban−Acree model[1−3] for a ternary solvent mixture

$$\ln S = \sum_{i=1}^{3} x_i \ln S_i + \sum_{1\le i<j\le 3} x_i x_j \left[\sum_{k=0}^{2} Q_k^{(ij)}(x_i - x_j)^k\right] \tag{16}$$

is commonly employed in solubility studies, where S and S_i are the solubilities of the solute in the mixture and in the ith pure solvent, respectively; $Q_k^{(ij)}$ are parameters calculated from particular physical properties of the solvents and the solute; and x_i is the fraction of ith solvent in the mixture. This model has been used for more than 100 solutes, and several solvents, like water, ethanol, propylene glycol, glycerin, glycofural, and polyethylene glycols, etc. Jouyban claimed that the model provides good accuracy compared to many other solubility models.

On the basis of the Jouyban−Acree model, an extended solvent mixture model was constructed as follows:

$$\ln S = \sum_{i=1}^{n} \alpha_i x_i + \sum_{1\le i<j\le n} x_i x_j \left[\sum_{k=0}^{2} \beta_k^{(ij)}(x_i - x_j)^k\right] + \gamma \prod_{i=1}^{n} (1 - x_i)^{r_i} \tag{17}$$

with unknown parameters α_i, $\beta_k^{(ij)}$, γ, and r_i. The parameters α_i, $\beta_k^{(ij)}$ are chosen in such a way that they have a similar magnitudes and sign as $\ln S_i$ and $Q_k^{(ij)}$ multiplied by a random number taken from $[0, 1]$. The last term is added to include high dimensional cooperative interactions among the solvents with r_i randomly taking values in $[0, 1]$. When $x_i = 1$, the last term vanishes because $x_i = 1$, which implies $x_j(j \neq i) = 0$ and no high dimensional cooperation exists. When $x_i = 0$ for some i's, we have $(1-x_i)^{r_i} = 1$, and cooperation among the remaining solvents is still included. The parameter γ is chosen to be 0.1,

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

0.5, and 1 of the standard deviation of ln S obtained by the Jouyban–Acree model, which implies that the contribution of the last term is correspondingly about 10, 50, and 100% of the contribution of the original terms given by Jouyban–Acree model to the output ln S. The model in eq 17 is strictly used as a basis to test the procedure in sections 2 and 3 in a multivariate correlated environment, while retaining a form related to the well-known Jouyban–Acree solvent model.

4.1. Ten Solvent Mixtures without Random Errors in the Data. One thousand data for solubility ln S were calculated from eq 17 for $n = 10$, where $\mathbf{x} = (x_1, ..., x_{10})$ were obtained by Dirichlet sampling with $\alpha_i = 1$, 0.1, 0.05 and biased random sampling, respectively. A separate set of 220 data points of solubility ln S for the {10, 3} simplex-lattice design on 1-, 2- and 3-dimensional boundaries (see Figure 6) were also generated for testing the boundary prediction of the HDMR model.

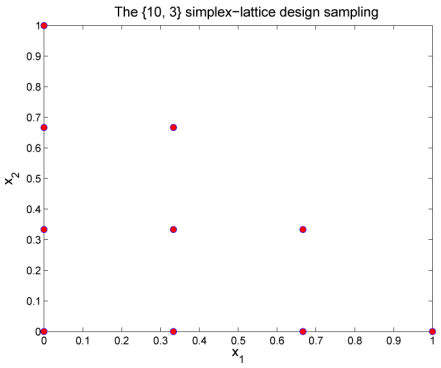


Figure 6. {10, 3} simplex-lattice design sampling projected into (x_1, x_2)-subspace. Projection into other 2-dimensional subspaces is the same.

Different numbers (80, 100, 120, 150, 200, 220) of data points, randomly selected from the 1000 data, were used to train the HDMR model, and the remaining data from the thousand samples were used for testing. The prediction accuracy was assessed as a criterion for the choice of whether to use second or third order HDMR expansions for different sample sizes of the training data. Finally, the resultant HDMR model was used to predict the 220 boundary points (boundary testing). As the performances for $\gamma = 0.1$, 0.5, and 1 are similar, only the results for $\gamma = 0.5$ are given in Table 1. For an assessment of the reported average absolute errors, a comparison can be made with the overall dynamic range of ln S in Figure 7.

For comparison, the results of the DoE {3, 3} simplex-lattice design for a three solvent mixture is also given in Table 1. In this case the 10 points of the {3, 3} simplex-lattice design (see Figure 1) were used as the training data, and the polynomial given in eq 3, was used as the model. The solubility of 46 randomly chosen points within the three variable simplex were calculated and used for testing the DoE design. Note that using the {3, 3} simplex-lattice design to generate all three solvent combinations from the 10 solvents is just the {10, 3} simplex-lattice design, which requires 220 data points. Figure 7 gives the truth plots of the HDMR models constructed from 120, 150, and 220 (all are not larger than 220) training data for full formulation testing and boundary testing.

We also tested the effect of all sampling schemes on the simple global polynomial model with $n = 10$ as employed in DoE (eq 3) and 175 unknown parameters

$$\ln S = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{1 \le i < j \le n} \beta_{ij} x_i x_j + \sum_{1 \le i < j < k \le n} \beta_{ijk} x_i x_j x_k \quad (18)$$

Table 1. Average Absolute Error of the 2nd or 3rd Order HDMR Models Constructed from the Solubility Data without Random Error

				average absolute error		
sampling method	data points of training	HDMR order	no. of unknowns	training	testing	boundary
Dirichlet	100	second	380	0.0000	0.0717	0.5941
$\alpha_i = 1$	150	third	3040	0.0000	0.1129	1.0775
	200	third	3040	0.0000	0.1255	1.8159
	220	third	3500	0.0000	0.1172	1.2012
Dirichlet	100	second	380	0.0000	0.1129	0.1746
$\alpha_i = 0.1$	150	third	3040	0.0000	0.0991	0.1439
	200	third	3040	0.0000	0.0564	0.0956
	220	third	3500	0.0000	0.0594	0.0983
Dirichlet	100	second	380	0.0000	2.0190	4.0883
$\alpha_i = 0.05$	150	third	3040	0.0044	2.6680	6.0354
	200	third	3040	0.0048	1.6182	3.9408
	220	third	3500	0.0044	3.2413	8.1300
biased random sampling	80	second	380	0.0000	0.1322	0.1756
	100	second	380	0.0000	0.0858	0.1156
	120	second	380	0.0000	0.0665	0.0919
	150	third	3500	0.0000	0.0722	0.0966
	200	third	3500	0.0000	0.0652	0.0946
	220	third	3500	0.0000	0.0688	0.0917
DoE	10[a]	eq 3	8[a]	0.0077	0.0980	–

[a]Refers to each three solvent mixture, a total of 220 data points for all cases.

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237–8249

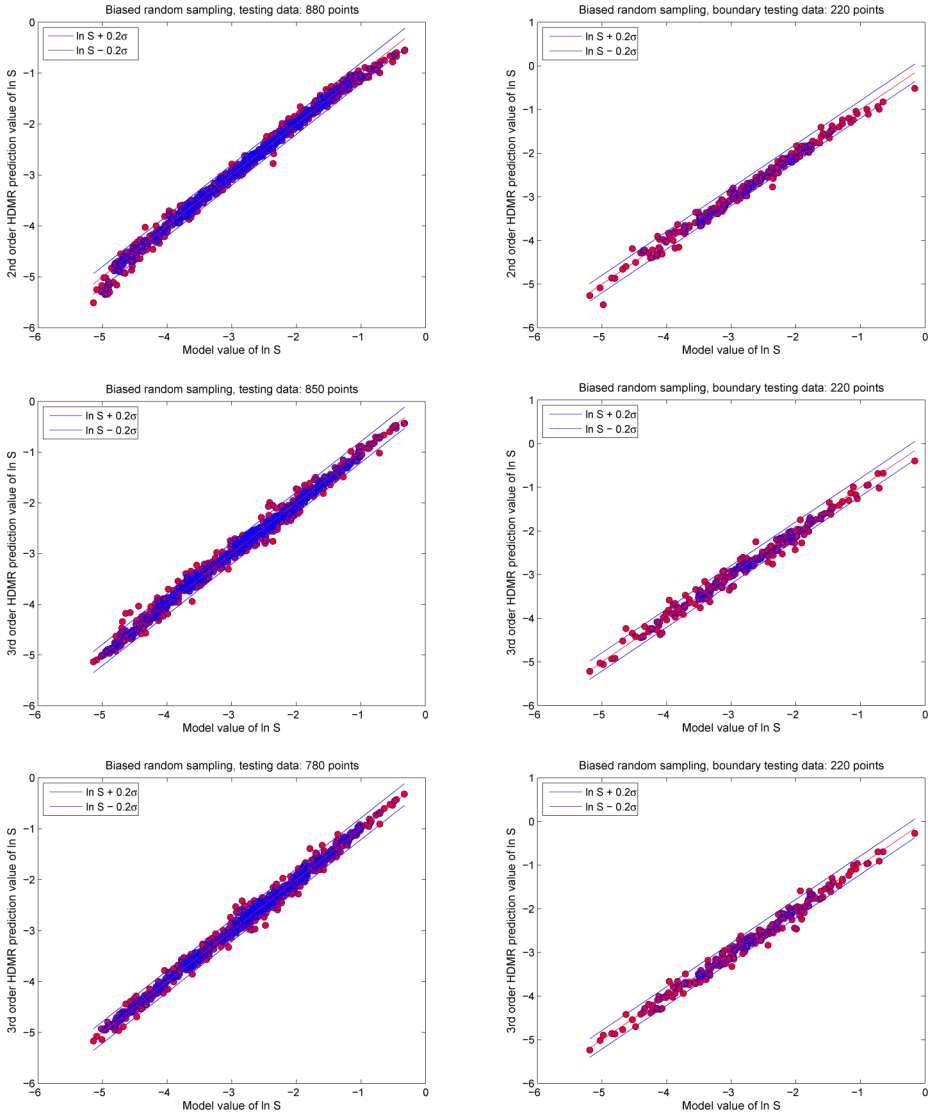


Figure 7. Truth plots for full formulation testing as well as boundary testing data. The second or third order HDMR models constructed from biased random sampling with 120 (upper panel), 150 (middle panel) and 220 (lower panel) training samples. Here, σ denotes the standard deviation of the ln S obtained from the 1000 data.

where β_0 is given by the average value of ln S for all training data. The other 175 coefficients $\{\beta\}$ are determined by least-squares regression. The results are given in Table 2.

From Tables 1, 2 and Figure 7 we may draw the following conclusions:

- The HDMR models constructed from Dirichlet sampling with α_i = 1 and 0.05 did not give satisfactory prediction of solubility on the 1-, 2-, and 3-dimensional boundaries, but the HDMR models constructed from both Dirichlet sampling with α_i = 0.1 and biased random sampling give good boundary predictions, and biased random sampling has the best performance. This result implies that sampling far away from the boundary or giving too much emphasis on the boundary can cause large errors of the resultant HDMR model for boundary prediction.
- The average absolute error of the DoE {3, 3} simplex-lattice design for the testing data is 0.098, which is very close to that of the HDMR models constructed from biased random sampling with 120 or more training data. For all possible {3, 3} simplex-lattice designs of 10 solvents, DoE requires 220 points. Therefore, under the same accuracy the biased random sampling combined with HDMR modeling shows a savings of ∼45% in the number of samples compared to that required by DoE. The sampling saving of biased random sampling with HDMR modeling is even larger when the number n of solvents is greater than 10.
- Note that in all cases of Table 1, the number of unknown parameters of the HDMR models is larger or *much* larger than the number of training data. D-MORPH regression is very effective under these circumstances, while other regression methods are often unable to handle this situation. This capability of D-MORPH regression is especially beneficial for practical applications where solvent preparation and testing is expensive.
- The HDMR modeling is also superior to the simple global polynomial model with all n variables. Since no hierarchical relation exists in eq 18, the D-MORPH regression cannot be applied and the least-squares regression has to be used to determine the coefficients

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

Table 2. Average Absolute Error of the 3rd Order Simple Polynomial Model Constructed from the Solubility Data without Random Error

sampling method	data points of training	average absolute error		
		training	testing	boundary
Dirichlet	100	0.0000	0.0464	0.2920
$\alpha_i = 1$	150	0.0000	0.0693	0.5445
	200	0.0062	0.0689	0.6399
	220	0.0085	0.0507	0.4445
Dirichlet	100	0.0000	0.2665	0.4147
$\alpha_i = 0.1$	150	0.0000	1.1052	1.8334
	200	0.0127	0.5124	0.8316
	220	0.0143	0.2267	0.3883
Dirichlet	100	0.0143	0.3652	0.7436
$\alpha_i = 0.05$	150	0.0207	51.410	117.18
	200	0.0186	381.05	747.02
	220	0.0217	236.31	732.94
biased random sampling	80	0.0000	0.0985	0.1465
	100	0.0000	0.1271	0.2019
	120	0.0000	0.1277	0.2066
	150	0.0000	0.2696	0.4574
	200	0.0098	0.2296	0.3820
	220	0.0167	0.2216	0.3485

$\{\beta\}$. When the coefficient matrix of the least-squares normal equation is singular, the solution of $\{\beta\}$ is given by the least-squares solution with the smallest L_2 norm, i.e., $A^+\mathbf{b}$ where A^+ is the generalized inverse of the coefficient matrix, and $\mathbf{b}$ is the right-hand side of the least-squares normal equation. This treatment is certainly worse than D-MORPH regression with singular value decomposition, and consequently has a larger error. Naturally, using eq 18, the effect of sampling schemes is different from that by using the HDMR model. One reason is that the coefficient matrix of the least-squares normal equation for eq 18 is still singular or very close to singular even if the number (e.g., 200, 220) of data is larger than the number of unknowns (175), especially for Dirichlet sampling with $\alpha_i = 0.1$ and 0.05. Table 2 shows that except of the data generated by Dirichlet sampling with $\alpha_i = 1$, the simple global polynomial model has larger errors for testing and boundary testing data compared to the HDMR model. Nevertheless, the biased random sampling still gives the best boundary prediction for eq 18.

4.2. Ten Solvent Mixtures with Random Errors in the Data. In practice, the measurements of solubility and the recorded solvent fractions have errors. To test the influence of the errors, 1000 data for solubility ln S with random error (the error has a normal distribution with zero mean and the standard deviation 0.1σ) were generated where $\mathbf{x} = (x_1, ..., x_{10})$ were obtained from biased random sampling. Similarly, a separate set of 220 data points for solubility ln S of the DoE {10, 3} simplex-lattice design were also considered with the same level of random error. The corresponding HDMR models constructed from 80, 100, 120, 150, 200, 220 points were used to test the remaining data of the 1000 points, and the 220 boundary data. The results are given in Table 3 and Figure 8.

Compared to Table 1, the accuracy of the HDMR models constructed from biased random sampling with random errors is a little worse, but the HDMR model is still better than DoE. The average absolute error of the {3, 3} simplex-lattice DoE design for the testing data is 0.1612, which is very close to that of the HDMR models constructed from biased random sampling with 120 or more training data. This test establishes the viability of the HDMR model with biased random sampling in the context of solubility, but we expect that this method will also apply to other classes of formulations.

5. CONCLUSIONS

Formulations or mixtures commonly arise in many applications. In some practical cases, a high dimensional choice of components may be considered, but the ultimate goal is to find the best mixture composition of, say two or three components, out of all possibilities. Finding suitable binary or ternary mixtures from all the permissible components, traditionally employs DoE, which normally uses the simplex-lattice design calling for performing many experiments when tens or more of permissible components are involved. This paper proposes a special experimental design to replace DoE simplex-lattice design based on (a) biased random sampling combined with (b) HDMR. Using the new method, the required number of experiments is significantly reduced, especially when the number of permissible components is large, while permitting satisfactory prediction for the full composition mixture as well as all possible binary and ternary mixtures. The procedure was tested with an extended solubility model modified from the Jouyban−Acree model with 10 solvents for illustration. A saving of ∼45% experiments was achieved for $n = 10$ solvents. The savings should rise with n because the number of experiments required by DoE simplex-lattice design grows very fast with n while the new method

Table 3. Average Absolute Error of the 2nd or 3rd Order HDMR Models Constructed from the Solubility Data with Random Error

sampling method	data points of training	HDMR order	number of unknown	average absolute error		
				training	testing	boundary
biased random sampling	80	second	380	0.0000	0.1624	0.2114
	100	second	380	0.0000	0.1388	0.1713
	120	third	3500	0.0000	0.1217	0.1549
	150	third	3500	0.0000	0.1176	0.1350
	200	third	3500	0.0000	0.1152	0.1475
	220	third	3500	0.0000	0.1153	0.1272
DoE	10[a]	eq 3	8[a]	0.0855	0.1612	–

[a]Refers to each three solvent mixture, a total of 220 data points for all case.

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

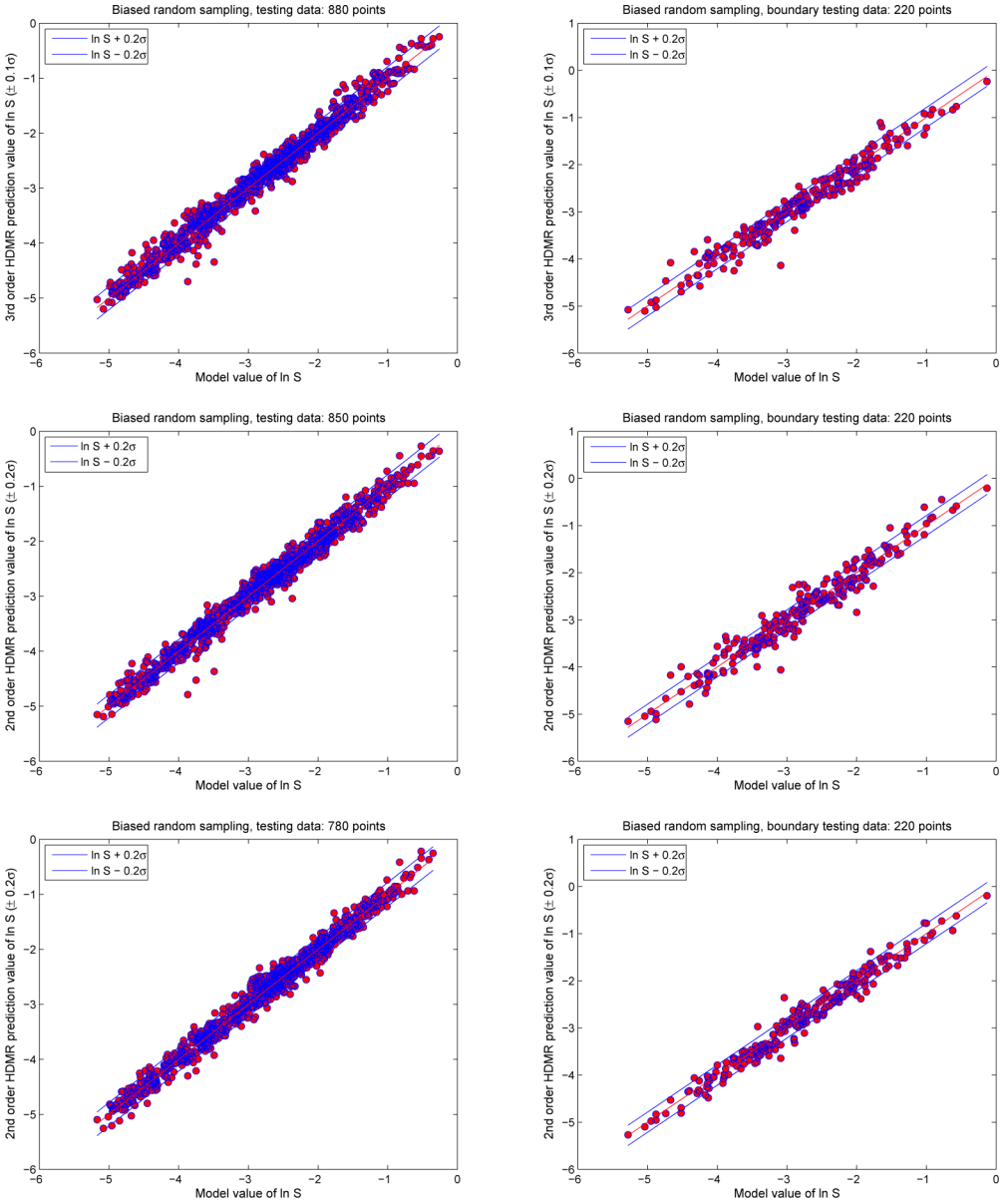


Figure 8. Truth plots for full formulation testing as well as boundary testing data. The second or third order HDMR models constructed from the solubility data with random errors. Biased random sampling with 120 (upper panel), 150 (middle panel), and 220 (lower panel) training samples were used, respectively.

introduced here scales much slower with the number of components.

Material formulations often consist of tens, but seldom hundreds of components. However, the HDMR modeling combined with D-MORPH regression has limitations when employed at very high dimensions.As shown in the key formula for the unknown parameters, eq 38, the algorithm needs to determine the generalized inverse A^{+} and perform singular value decomposition of matrix PB. Both have a dimension equal to the number of unknowns, which can be 3500 for the example with $n = 10$ used in this paper. The number of unknowns will be even larger for larger n, and it may be difficult to treat it. To solve this problem, we have developed a new approach, which combines support vector regression with HDMR. In this method the basis function expansion is replaced by kernels to avoid high dimensionality. We have tested a solvent mixture screening model with 50 solvents, and 200 biased random samples are sufficient to construct a satisfactory prediction model. This new development will be reported later in another paper.[20,21]

APPENDIX

HDMR Methodology

The HDMR component functions $f_u(\mathbf{x}_u)$ with independent and/or correlated input variables are constructed by a method: the combination of extended bases and D-MORPH regression. The details of the HDMR methodology are given below.

A. Extended Bases.[11,17] The analytical form of HDMR component functions may not be obtained for an arbitrary function $f(\mathbf{x})$ with an arbitrary probability distribution of $\mathbf{x}$. However, in practice, the HDMR component function may be approximate by a combination of some suitable basis functions. The *sufficient* condition for *hierarchical orthogonality* of the component functions is that the subspace of the Hilbert space

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

spanned by the basis functions for any lower order component function is a normal subspace of the subspace spanned by the basis functions of their nested higher order component functions. Suppose that a subspace V in Hilbert space is spanned by the basis $\{v_1, v_2, ..., v_k\}$, and a larger subspace U ($\supset V$) is spanned by the extended basis $\{v_1, v_2, ..., v_k, v_{k+1}, ..., v_m\}$. Then U can be decomposed as

$$U = V \oplus V^{\perp}$$

where $V^{\perp}$ is the orthogonal complementary subspace of V in U. One can always find a vector in $V^{\perp}$ (i.e., a certain linear combination of $v_1, v_2, ..., v_k, v_{k+1}, ..., v_m$) orthogonal to all vectors in V.

To satisfy this sufficient condition, the component functions are approximated by expansions in some suitable basis functions $\{\varphi\}$ (polynomials, splines, etc.) as follows

$$f_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^{(0)i} \varphi_r^i(x_i) \tag{19}$$

$$f_{ij}(x_i, x_j) \approx \sum_{r=1}^{k} [\alpha_r^{(ij)i} \varphi_r^i(x_i) + \alpha_r^{(ij)j} \varphi_r^j(x_j)] + \sum_{p=1}^{l} \sum_{q=1}^{l} \beta_{pq}^{(0)ij} \varphi_p^i(x_i) \varphi_q^j(x_j), \tag{20}$$

$$f_{ijk}(x_i, x_j, x_k) \approx \sum_{r=1}^{k} [\alpha_r^{(ijk)i} \varphi_r^i(x_i) + \alpha_r^{(ijk)j} \varphi_r^j(x_j) + \alpha_r^{(ijk)k} \varphi_r^k(x_k)] + \sum_{p=1}^{l} \sum_{q=1}^{l} [\beta_{pq}^{(ijk)ij} \varphi_p^i(x_i) \varphi_q^j(x_j) + \beta_{pq}^{(ijk)ik} \varphi_p^i(x_i) \varphi_q^k(x_k) + \beta_{pq}^{(ijk)jk} \varphi_p^j(x_j) \varphi_q^k(x_k)] + \sum_{p=1}^{m} \sum_{q=1}^{m} \sum_{r=1}^{m} \gamma_{pqr}^{(0)ijk} \varphi_p^i(x_i) \varphi_q^j(x_j) \varphi_r^k(x_k), \cdots \tag{21}$$

where k, l, m are integers. Note that in eqs 19-21 the basis functions of the lower order component functions are always a subset of those for the higher order ones, and the hierarchical orthogonality between $f_i(x_i)$, $f_{ij}(x_i, x_j)$, $f_{ijk}(x_i, x_j, x_k)$, ... can be readily achieved. Consider an example, $f_{ij}(x_i, x_j)$. One can always find suitable values for $\alpha_r^{(ij)i}$, $\alpha_r^{(ij)j}$ and $\beta_{pq}^{(0)ij}$ such that $f_{ij}(x_i, x_j)$ is orthogonal to $\varphi_r^i(x_i)$, $\varphi_r^j(x_j)$, as well as any linear combinations of them, and consequently, $f_i(x_i)$ and $f_j(x_j)$ as demanded by eq 13.

The optimal orthonormal polynomials $\{\varphi\}$ satisfying the conditions

$$\int w_i(x_i) \varphi_r^i(x_i)\, dx_i \approx \sum_{s=1}^{N} \varphi_r^i(x_i^{(s)})/N = 0, \quad \text{for all } r, i \tag{22}$$

$$\int w_i(x_i) (\varphi_r^i(x_i))^2\, dx_i \approx \sum_{s=1}^{N} (\varphi_r^i(x_i^{(s)}))^2/N = 1, \quad \text{for all } r, i \tag{23}$$

$$\int w_i(x_i) \varphi_p^i(x_i) \varphi_q^i(x_i)\, dx_i \approx \sum_{s=1}^{N} \varphi_p^i(x_i^{(s)}) \varphi_q^i(x_i^{(s)})/N = 0, \quad p \neq q \tag{24}$$

where $w_i(x_i)$ is the marginal probability density function (pdf) of x_i, and their tensor products are used to construct the basis functions. *For the sake of notational neatness, we omit the specific integration dimension and range and use $\int$ to represent all integrations in this paper.*

In this fashion the bases may be constructed from a set of data generated according to a given pdf $w(\mathbf{x})$, where $x_i^{(s)}$ is the sth sample and N is the total number of samples. The basis set members have zero mean, unit norm and are mutually orthogonal with respect to the marginal pdf weight $w_i(x_i)$. In many cases, satisfactory accuracy is likely attainable using only $\varphi_r^i(x_i)$, $r \leq 3$ to approximate $f_i(x_i)$, $f_{ij}(x_i, x_j)$ and $f_{ijk}(x_i, x_j, x_k)$.

Using eqs 19−21, the third order HDMR expansion for an n-variate function $f(\mathbf{x})$ can be expressed as

$$f(\mathbf{x}) \approx f_0 + \sum_{i=1}^{n} \sum_{r=1}^{k} (\alpha_r^{(0)i} + \sum_{j=1/j\neq i}^{n} \alpha_r^{(ij)i} + \sum_{j<k=1/j,k\neq i}^{n} \alpha_r^{(ijk)i}) \varphi_r^i(x_i) + \sum_{1\leq i<j\leq n} \sum_{p=1}^{l} \sum_{q=1}^{l} (\beta_{pq}^{(0)ij} + \sum_{k=1/k\neq i,j}^{n} \beta_{pq}^{(ijk)ij}) \varphi_p^i(x_i) \varphi_q^j(x_j) + \sum_{1\leq i<j<k\leq n} \sum_{p=1}^{m} \sum_{q=1}^{m} \sum_{r=1}^{m} \gamma_{pqr}^{(0)ijk} \varphi_p^i(x_i) \varphi_q^j(x_j) \varphi_r^k(x_k) \tag{25}$$

The value of f_0 is estimated by

$$f_0 = \frac{1}{N} \sum_{s=1}^{N} f(\mathbf{x}^{(s)}) \tag{26}$$

The constant coefficients $\{\alpha\}$, $\{\beta\}$, and $\{\gamma\}$ are unknown parameters. Consider some examples. For the second order HDMR expansion with $n = 10$, $k = l = 2$, the total number of unknown parameters is 380. For the third order HDMR expansion with $n = 10$, $k = 1$, $l = m = 2$, the total number of unknown parameters is 3040. For the third order HDMR expansion with $n = 10$, $k = l = m = 2$, the total number of unknown parameters is 3500. These unknown parameters will be determined by D-MORPH regression (see Appendix B).

Equation 25 can be written in a vector form for all of the data

$$\phi(\mathbf{x}^{(s)})^T \mathbf{c} = f(\mathbf{x}^{(s)}) - f_0, \quad (s = 1, 2, ..., N) \tag{27}$$

Here vector $\mathbf{c}$ is composed of all the unknown parameters, and vector $\phi(\mathbf{x}^{(s)})^T$ consists of the values of the corresponding basis functions at $\mathbf{x}^{(s)}$.

Equation 27 can be written in matrix form as

$$\mathbf{\Phi c} = \mathbf{b} \tag{28}$$

where Φ is an $N \times t$ matrix (the sth row of Φ is $\phi(\mathbf{x}^{(s)})^T$), t is total number of unknown parameters, and $\mathbf{b}$ is an N-dimensional vector whose sth element is $f(\mathbf{x}^{(s)}) - f_0$. Because of the extended bases, $\phi(\mathbf{x}^{(s)})^T$ has repeated elements, some columns of Φ are identical.

The normal equation of eq 28 for the least-squares regression is

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

$$\frac{1}{N}\Phi^T\Phi\mathbf{c} = \frac{1}{N}\Phi^T\mathbf{b} \tag{29}$$

As Φ^T has duplicate rows, some equations in eq 29 are identical. These duplicate equations are redundant and can be removed. The resultant linear algebraic equation system is

$$A\mathbf{c} = \mathbf{d} \tag{30}$$

where A and $\mathbf{d}$ are just $\Phi^T\Phi/N$ and $\Phi^T\mathbf{b}/N$ after removing the duplicate rows. Now A is a $p \times t (p < t)$ rectangular matrix. In eq 30 the number (t) of unknown parameters is larger than the number (p) of equations. Such a system is consistent and has an infinite number of solutions for $\mathbf{c}$ with the general form

$$\mathbf{c} = A^+\mathbf{d} + (\mathrm{I}_t - A^+A)\mathbf{u} \tag{31}$$

where I_t is the identity matrix with dimension t and $\mathbf{u}$ is an arbitrary vector in $\mathbb{R}^t$, and A^+ is the generalized inverse of A satisfying all four Penrose conditions.[20]

B. D-MORPH Regression.[18,19] The HDMR development in Appendix A led to an underdetermined consistent linear algebraic equation system, which has an infinite number of solutions composing a completely connected submanifold $\mathcal{M}$. D-MORPH regression is a practical means to search for a solution within $\mathcal{M}$ satisfying an extra requirement of minimizing a chosen cost function $\mathcal{K}$. Equation 30 is such a system and the hierarchical orthogonality of the component functions is the extra requirement. D-MORPH regression searches for a solution satisfying an extra requirement by considering an exploration path $\mathbf{c}(s)$ within $\mathcal{M}$ with s in $[0, \infty)$, which satisfies a differential equation obtained by differentiation of eq 31 with respect to s

$$\begin{aligned}\frac{\mathrm{d}\mathbf{c}(s)}{\mathrm{d}s} &= (\mathrm{I}_t - A^+A)\frac{\mathrm{d}\mathbf{u}(s)}{\mathrm{d}s}\\ &= (\mathrm{I}_t - A^+A)\mathbf{v}(s) = P\mathbf{v}(s),\end{aligned} \tag{32}$$

where P is an orthogonal projector satisfying

$$P^2 = P, \quad P^T = P \tag{33}$$

which yields

$$P = P^2 = P^TP \tag{34}$$

The function vector $\mathbf{v}(s)$ may be freely chosen to not only enable broad choices for exploring $\mathbf{c}(s)$, but to also continuously reduce a defined cost $\mathcal{K}(\mathbf{c}(s))$ (e.g., the model variance, fitting smoothness, the weighted norm of $\mathbf{c}$, or particularly here the hierarchical orthogonality of the component functions) along the exploration path. If the free function vector is chosen as

$$\mathbf{v}(s) = -\frac{\partial\mathcal{K}(\mathbf{c}(s))}{\partial\mathbf{c}} \tag{35}$$

then we obtain

$$\begin{aligned}\frac{\mathrm{d}\mathcal{K}(\mathbf{c}(s))}{\mathrm{d}s} &= \left(\frac{\partial\mathcal{K}(\mathbf{c}(s))}{\partial\mathbf{c}}\right)^T\frac{\mathrm{d}\mathbf{c}(s)}{\mathrm{d}s} = \left(\frac{\partial\mathcal{K}(\mathbf{c}(s))}{\partial\mathbf{c}}\right)^T P\mathbf{v}(s)\\ &= -\left(P\frac{\partial\mathcal{K}(\mathbf{c}(s))}{\partial\mathbf{c}}\right)^T\left(P\frac{\partial\mathcal{K}(\mathbf{c}(s))}{\partial\mathbf{c}}\right) \le 0\end{aligned} \tag{36}$$

i.e., the cost $\mathcal{K}$, used as an additional requirement, will be continuously reduced (systematically refining the model) over the course of traversing $s \ge 0$. Therefore,

$$\mathbf{c}_\infty = \lim_{s\to\infty}\mathbf{c}(s)$$

is the solution which minimizes $\mathcal{K}$. When the cost function is defined as a quadratic form in $\mathbf{c}$

$$\mathcal{K} = \frac{1}{2}\mathbf{c}^TB\mathbf{c} \tag{37}$$

where B is symmetric and non-negative definite, the analytical form of $\mathbf{c}_\infty$ has been obtained as

$$\mathbf{c}_\infty = V_{t-r}(U_{t-r}^TV_{t-r})^{-1}U_{t-r}^TA^+\mathbf{d} \tag{38}$$

where U_{t-r}, and V_{t-r} are the last $t-r$ columns of U and V obtained by singular value decomposition of PB [21]

$$PB = U\begin{bmatrix}S_r & 0\\ 0 & 0\end{bmatrix}V^T \tag{39}$$

Equation 38 is the key practical formula for the optimal solution $\mathbf{c}$ obtained by D-MORPH regression. This solution $\mathbf{c}_\infty$ is unique in $\mathcal{M}$ corresponding to the global minimum of the cost function. The new solution $\mathbf{c}_\infty$ given by D-MORPH regression is simply a linear combination of the elements of $\mathbf{c}$ obtained by least-squares regression with the minimum of $\| \mathbf{c} \|_{l_2}$ (i.e., $A^+\mathbf{d}$).

C. Construction of Cost Function.[11] The solution of eq 38 satisfying the hierarchical orthogonality condition can be determined by constructing a proper cost function, i.e., the symmetric matrix B.

The first order component function, $f_i(x_i)$, is required to be orthogonal to the zeroth order component function, f_0, i.e.,

$$\int f_0f_i(x_i)w_i(x_i)\,\mathrm{d}x_i = f_0\int f_i(x_i)w_i(x_i)\,\mathrm{d}x_i = 0, \quad (i = 1, 2, ..., n) \tag{40}$$

Since f_0 may be nonzero, the necessary and sufficient condition for eq 40 is

$$\int f_i(x_i)w_i(x_i)\mathrm{d}x_i = \mathbb{E}[f_i(x_i)] = 0, \quad (i = 1, 2, ..., n) \tag{41}$$

When $f_i(x_i)$ is represented as a linear combination of basis functions $\varphi_j(x_i)$ $(j = q + 1, ..., q + k)$, we have

$$\int\sum_{j=q+1}^{q+k}c_j\varphi_j(x_i)w_i(x_i)\,\mathrm{d}x_i \approx \sum_{j=q+1}^{q+k}c_j\Big(\sum_{s=1}^{N}\varphi_j(x_i^{(s)})/N\Big) = 0 \tag{42}$$

Equation 42 can be written as

$$\begin{aligned}&\Big(\sum_{s=1}^{N}\varphi_{q+1}(x_i^{(s)})/N \quad \sum_{s=1}^{N}\varphi_{q+2}(x_i^{(s)})/N \quad \cdots \quad \sum_{s=1}^{N}\varphi_{q+k}(x_i^{(s)})/N\Big)\\ &\begin{pmatrix}c_{q+1}\\ c_{q+2}\\ \vdots\\ c_{q+k}\end{pmatrix} = 0\end{aligned} \tag{43}$$

or in vector form

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

$$\mathbf{Sr}(x_i)^T \mathbf{c}^i = 0, \quad (i = 1, 2, ..., n) \tag{44}$$

Here, $\mathbf{c}^i$ is the vector composed of all unknown parameters related to $f_i(x_i)$.

This is a scalar equation, and the corresponding cost function for $f_i(x_i)$ is set to be

$$\mathcal{K}^i = \frac{1}{2}(\mathbf{c}^i)^T \mathbf{Sr}(x_i)\mathbf{Sr}(x_i)^T \mathbf{c}^i = \frac{1}{2}(\mathbf{c}^i)^T B^i \mathbf{c}^i \quad (i = 1, 2, ..., n) \tag{45}$$

where B^i is a $k \times k$ symmetric and non-negative definite matrix. Therefore, $\mathcal{K}^i \geq 0$ with the minimum value being zero. $\mathcal{K}^i$ is zero if and only if $\mathbf{Sr}(x_i)^T\mathbf{c}^i$ is zero, i.e., eq 44 (consequently, eq 41) is satisfied. When optimal orthonormal polynomials are used as $\varphi_j(x_i)$, then all of the sums $\sum_{s=1}^{N} \varphi_{q+j}(x_i^{(s)})/N (j = 1, 2, 3)$ are zero (see eqs 22−24). In this circumstance B^i is a null matrix, which implies that there is no need for further restriction on the expansion coefficients for $f_i(x_i)$ upon using optimal orthonormal polynomials.

The second order component function $f_{ij}(x_i, x_j)$ is required to be orthogonal to f_0 and the first order component functions, $f_i(x_i)$ and $f_j(x_j)$. This can be achieved by setting $f_{ij}(x_i, x_j)$ to be orthogonal to all the basis functions used in f_0 (its basis is 1), $f_i(x_i)$ and $f_j(x_j)$. Since $f_{ij}(x_i, x_j)$ is orthogonal to all the basis functions, it must be orthogonal to any linear combination of these basis functions, and consequently orthogonal to f_0, $f_i(x_i)$ and $f_j(x_j)$.

Let

$$f_i(x_i) = \sum_{l=1}^{k} c_l^i \varphi_l^i(x_i) \tag{46}$$

$$f_j(x_j) = \sum_{l=1}^{k} c_l^j \varphi_l^j(x_j) \tag{47}$$

$$f_{ij}(x_i, x_j) = \sum_{l=1}^{k} c_l^{(ij)i} \varphi_l^i(x_i) + \sum_{l=1}^{k} c_l^{(ij)j} \varphi_l^j(x_j) + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij} \varphi_p^i(x_i)\varphi_q^j(x_j) \tag{48}$$

The orthogonality between $f_{ij}(x_i, x_j)$ and f_0 is given by

$$\begin{aligned}
&\int 1\Big(\sum_{l=1}^{k} c_l^{(ij)i} \varphi_l^i(x_i) + \sum_{l=1}^{k} c_l^{(ij)j} \varphi_l^j(x_j) \\
&\quad + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij} \varphi_p^i(x_i)\varphi_q^j(x_j)\Big) w_{ij}(x_i, x_j)\, dx_i\, dx_j \\
&\approx \sum_{l=1}^{k} c_l^{(ij)i}\Big(\sum_{s=1}^{N} \varphi_l^i(x_i^{(s)})/N\Big) + \sum_{l=1}^{k} c_l^{(ij)j}\Big(\sum_{s=1}^{N} \varphi_l^j(x_j^{(s)})/N\Big) \\
&\quad + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij}\Big(\sum_{s=1}^{N} \varphi_p^i(x_i^{(s)})\varphi_q^j(x_j^{(s)})/N\Big) \\
&\quad = \mathbf{Sr}_0(x_i, x_j)^T \mathbf{c}^{ij} = 0
\end{aligned} \tag{49}$$

where $\mathbf{c}^{ij}$ is a t_{ij} $(=2k + (l')^2)$-dimensional vector consisting of all expansion coefficients for $f_{ij}(x_i, x_j)$.

The orthogonality between $f_{ij}(x_i, x_j)$ and the basis $\varphi_v^i(x_i)$ is given by

$$\begin{aligned}
&\int \varphi_v^i(x_i)\Big(\sum_{l=1}^{k} c_l^{(ij)i} \varphi_l^i(x_i) + \sum_{l=1}^{k} c_l^{(ij)j} \varphi_l^j(x_j) \\
&\quad + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij} \varphi_p^i(x_i)\varphi_q^j(x_j)\Big) w_{ij}(x_i, x_j)\, dx_i dx_j \\
&= \sum_{l=1}^{k} c_l^{(ij)i} \langle \varphi_v^i(x_i), \varphi_l^i(x_i)\rangle + \sum_{l=1}^{k} c_l^{(ij)j} \langle \varphi_v^i(x_i), \varphi_l^j(x_j)\rangle \\
&\quad + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij} \langle \varphi_v^i(x_i), \varphi_p^i(x_i)\varphi_q^j(x_j)\rangle \\
&\approx \sum_{l=1}^{k} c_l^{(ij)i}\Big(\sum_{s=1}^{N} \varphi_v^i(x_i^{(s)})\varphi_l^i(x_i^{(s)})/N\Big) \\
&\quad + \sum_{l=1}^{k} c_l^{(ij)j}\Big(\sum_{s=1}^{N} \varphi_v^i(x_i^{(s)})\varphi_l^j(x_j^{(s)})/N\Big) \\
&\quad + \sum_{p=1}^{l'}\sum_{q=1}^{l'} c_{pq}^{(0)ij}\Big(\sum_{s=1}^{N} \varphi_v^i(x_i^{(s)})\varphi_p^i(x_i^{(s)})\varphi_q^j(x_j^{(s)})/N\Big) \\
&= \mathbf{Sr}_{iv}(x_i, x_j)^T \mathbf{c}^{ij} = 0, \quad (v = 1, 2, ..., k)
\end{aligned} \tag{50}$$

where the elements of $\mathbf{Sr}_{iv}(x_i, x_j)^T$ are the estimates of the inner products of $\varphi_v^i(x_i)$ and all the basis functions used by $f_{ij}(x_i, x_j)$. The orthogonality between $\varphi_v^j(x_j)$ and $f_{ij}(x_i, x_j)$ can be treated similarly, which gives

$$\mathbf{Sr}_{jv}(x_i, x_j)^T \mathbf{c}^{ij} = 0, \quad (v = 1, 2, ..., k) \tag{51}$$

All together there are $2k + 1$ equations in eqs 49−51, which can be represented in matrix form

$$\mathbf{Sr}(x_i, x_j)^T \mathbf{c}^{ij} = \mathbf{0} \tag{52}$$

where $\mathbf{Sr}(x_i, x_j)^T$ is a $(2k + 1) \times t_{ij}$ matrix, and $\mathbf{0}$ is a $(2k + 1)$-dimensional null vector.

The cost function for the orthogonality between $f_{ij}(x_i, x_j)$ and f_0, $f_i(x_i)$, $f_j(x_j)$ is specified as

$$\mathcal{K}^{ij} = \frac{1}{2}(\mathbf{c}^{ij})^T \mathbf{Sr}(x_i, x_j)\mathbf{Sr}(x_i, x_j)^T \mathbf{c}^{ij} = \frac{1}{2}(\mathbf{c}^{ij})^T B^{ij} \mathbf{c}^{ij}, \quad (i < j = 1, 2, ..., n) \tag{53}$$

where B^{ij} is a $t_{ij} \times t_{ij}$ symmetric, non-negative definite matrix. Therefore, $\mathcal{K}^{ij} \geq 0$ with a minimum value of zero which occurs if and only if $\mathbf{Sr}(x_i, x_j)^T\mathbf{c}^{ij}$ is a null vector, i.e., $f_{ij}(x_i, x_j)$ is orthogonal to f_0 and all $\varphi_v^i(x_i)$ and $\varphi_v^j(x_j)$.

A similar treatment can be made for $f_{ijk}(x_i, x_j, x_k)$, and the corresponding cost function

$$\mathcal{K}^{ijk} = \frac{1}{2}(\mathbf{c}^{ijk})^T B^{ijk} \mathbf{c}^{ijk}, \quad (i < j < k = 1, 2, ..., n) \tag{54}$$

can be constructed. If the third order HDMR expansion is used, the total cost function is set to be

$$\mathcal{K} = \sum_{i=1}^{n} \mathcal{K}^i + \sum_{1 \leq i < j \leq n} \mathcal{K}^{ij} + \sum_{1 \leq i < j < k \leq n} \mathcal{K}^{ijk} = \frac{1}{2}\mathbf{c}^T B \mathbf{c} \tag{55}$$

where $\mathbf{c}$ consists of all the unknown coefficients in eq 25, and

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249

$$B = \begin{pmatrix} B^{1} & & & & & & & & \\ & \ddots & & & & & & & \\ & & B^{n} & & & & & & \\ & & & B^{12} & & & & & \\ & & & & \ddots & & & & \\ & & & & & B^{(n-1)n} & & & \\ & & & & & & B^{123} & & \\ & & & & & & & \ddots & \\ & & & & & & & & B^{(n-2)(n-1)n} \end{pmatrix} \tag{56}$$

is a non-negative definite matrix. Therefore, $\mathcal{K} \geq 0$ and its minimum value is zero which implies the hierarchical orthogonality of the component functions. All B^{ij} and B^{ijk} are submatrices of $\Phi^T\Phi/N$ and can be obtained from it.

AUTHOR INFORMATION

Corresponding Author

*(H.R.) Telephone: 609-258-3917. E-mail: hrabitz@princeton.edu.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Support of this work is provided by ONR with Account Number N00014-11-1-0716.

REFERENCES

(1) Jouyban, A.; et al. Solubility Prediction of Paracetamol in Binary and Ternary Solvent Mixtures Using Jouyban-Acree Model. *Chem. Pharm. Bull.* **2006**, *54* (4), 428−431.
(2) Jouyban, A. Review of the Cosolvency Models for Predicting Solubility of Drugs in Water-Cosolvent Mixtures. *J. Pharm. Pharmaceut. Sci.* **2008**, *11* (1), 32−58.
(3) Jouyban, A.; et al. Solubility Prediction of Drugs in Mixed Solvents Using Partial Solubility Parameters. *J. Pharm. Sci.* **2011**, *100* (10), 4368−4382.
(4) NIST, *Engineering Statistics*, www.itl.nist.gov/div898/handbook/pri/section5/pri542.htm.
(5) Smith, N. A., Tromble, R. W. *Sampling Uniformly from the Unit Simplex, Department of Computer Science/Center for Language and Speech Processing*; Johns Hopkins University: Baltimore, MD, 2004.
(6) Fieldsend, J. E. *A Short Note on the Efficient Random Sampling of the Multi-Dimensional Pyramid between a Simplex and the Origin Lying in the Unit, School of Engineering, Computer Science and Mathematics*, University of Exeter: Exeter, U.K. 24th August 2005.
(7) Hazewinkel, M., Ed. Dirichlet Distribution. In *Encyclopedia of Mathematics*; Springer: Berlin, 2001.
(8) Rabitz, H.; Alis, O. F. General Foundations of High-Dimensional Model Representations. *J. Math. Chem.* **1999**, *25*, 197−233.
(9) Alis, O. F.; Rabitz, H. Efficient Implementation of High Dimensional Model Representations. *J. Math. Chem.* **2001**, *29* (2), 127−142.
(10) Li, G.; Rosenthal, C.; Rabitz, H. High-Dimensional Model Representations. *J. Phys. Chem. A* **2001**, *105* (33), 7765−7777.
(11) Li, G.; Rabitz, H. General Formulation of HDMR Component Functions with Independent and Correlated Variables. *J. Math. Chem.* **2012**, *50*, 99−130.
(12) Li, G.; Rabitz, H. Analytical HDMR Formulas for Functions Expressed as Quadratic Polynomials with a Multivariate Normal Distribution. *J. Math. Chem.* **2014**, *52*, 2052−2073.
(13) Hoeffding, H. A Class of Statistics with Asymptotically Normal Distribution. *Ann. Math. Stat.* **1948**, *19* (3), 293−325.
(14) Sobol, M. Sensitivity Estimates for Nonlinear Mathematical Models. *Math. Model.* **1990**, *2*, 112−118; in Russian; translated in *Mathematical Modelling and Computational Experiments* **1993**, *1*, 407−414.
(15) Dunkl, C., Xu, Y. *Orthogonal polynomials of several variables*; Cambridge University Press: Cambridge, U.K, 2001.
(16) Rahman, S. A Generalized ANOVA Dimensional Decomposition for Dependent Probability Measures. *SIAM/ASA J. Uncertainty Quant.* **2014**, *2* (1), 670−697.
(17) Li, G.; et al. Random Sampling-High Dimensional Model Representation (RS-HDMR) and Orthogonality of Its Different Order Component Functions. *J. Phys. Chem. A* **2006**, *110*, 2474−2485.
(18) Li, G.; Rabitz, H. D-MORPH Regression: Application to Modeling with Unknown Parameters More than Observation Data. *J. Math. Chem.* **2010**, *48*, 1010−1035.
(19) Li, G.; Rey-de-Castro, R.; Rabitz, H. D-MORPH Regression for Modeling with Fewer Unknown Parameters than Observation Data. *J. Math. Chem.* **2012**, *50*, 1747−1764.
(20) Rao, C. R.; Mitra, S. K. *Generalized Inverse of Matrix and its Applications*; Wiley: New York, 1971.
(21) Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. *Numerical Recipes in FORTRAN−The Art of Science Computing*, 2nd ed.,Cambridge University Press: New York, 1992; p 51.

DOI: 10.1021/acs.jpca.5b04911
J. Phys. Chem. A 2015, 119, 8237−8249