Model Selection for Analysis of Uniform Design and Computer Experiment *

Runze Li
Department of Statistics
Pennsylvania State University
University Park, PA 16802-2111

Abstract

In this paper, a new variable selection procedure is introduced for the analysis of uniform design and computer experiment. The new procedure is distinguished from the traditional ones in such a way that it deletes insignificant variables and estimates the coefficients of significant variables simultaneously. The new procedure has an oracle property (Fan and Li [8]). It is better than the best subset variable selection in terms of computational cost and model stability. It is superior to the stepwise regression because it does not ignore stochastic errors during the course of selecting variables. The proposed procedure is illustrated by two examples, one is a typical example of uniform design, and the other one is a classical example for computer experiment.

Keywords: Computer Experiment; Uniform Design; Variable Selection.

1 Introduction

Experimental designs are very useful for screening active-effect factors among a large number of potential factors in various scientific research and industrial management. Many experimental designs have been proposed to handle practical problems. Among various experimental designs, orthogonal designs have been used in agricultural as well as industrial problems since 1920s. Modern scientific researches impose many challenging tasks for experimental designs. Uniform design (Fang [9]) seeks its design points to be uniformly scattered on the experimental domain, and is becoming popular since 1980. Many scientific phenomena may be studied by computer simulation. To save times and cost, computer experimental designs play a crucial role in computer simulation studies. Unlike analysis of orthogonal designs and optimal designs, analysis of uniform design and computer experiment are much more challenging. Variable selection is fundamental in the stage of analyzing uniform design and computer experimental design. Hence, the goal of this paper is to introduce some new methodology of model selection developed in the very recent statistical literature.

Fang [9] and Wang and Fang [27] proposed the uniform design to handle a practical design problem with 5 factors, each factor having 31 levels, and with requirement that the number of experiments

* This research was supported by a NSF grant DMS-0102505.

must be less than 50. This challenging design problem was solved by using a uniform design with 31 experimental times. The uniform design enjoys several advantages. The uniform design retains the uniform property of fractional factorial design, but it dramatically reduces the number of experimental times required for the factorial designs when the numbers of levels of some potential factors are large. Compared with optimal designs, the uniform design enjoys its robustness since it does not require to specify a statistical model before experimenters conduct their experiments. Therefore experimenters can use uniform design to explore relationships between the response and the potential candidate factors with a reasonable number of experimental times and without model pre-specification. Due to its advantages, the uniform design has been very popular in China and becoming popular in various research fields involved experimental designs. The analysis of uniform design, however, is challenging due to its small number of observations and because experimenters do not specify a statistical model in advance. Hence, analysis of uniform design is crucial after conducting experiments.

Computer simulation becomes widely used in science and engineering to investigate complicated physical phenomena. Output obtained from a computer experiment is deterministic. This imposes challenge in analyzing such data. Many complex computer models have been proposed in the statistical literature. Sack, Welch, Mitchell and Wynn [24] gives a comprehensive review on this topic. They suggest to model the deterministic output as a realization of a stochastic process, and employ Bayesian kriging models, consisting of two components: a general linear model and a realization of a stationary Guassian random process. Koehler and Owen [18] provides a detailed review on how to scatter computer design points over the experimental domain effectively and how to analyze the deterministic output by employing statistical tools in the analysis of linear regression. It has been recognized that uniform design is very useful for constructing of effective computer experiments. In addition, model selection plays an important role in the analysis of computer experiment.

In the initial stage of analysis of uniform design and computer experiments, many predictor variables, such as the linear, quadratic and interaction terms associated with candidate factors, are introduced to a linear regression model in order to attenuate modeling bias. On the other hand, the goal of experiment is to obtain a model with good prediction power. Thus, to enhance model predictibility, variable selection plays important role in the analysis of uniform design and computer experimental design. Many variable selection criteria have been developed in statistical literature. To implement these criteria, data analysts usually use stepwise regression procedure or the best subset variable selection to find a good subset. However, the best subset variable selection suffers from several drawbacks, the most severe one of which is its lack of stability, as analyzed, for instance, by Breiman [5]. The stepwise deletion procedures ignore stochastic errors inherent in the process of selecting variables. Therefore, the sampling properties of the resulting estimates are difficult to understand. In an attempt of automatically selecting significant variable, Fan and Li [8] proposed a family of variable selection procedures via penalized likelihood approach. Their approach is distinguished from the traditional ones in that the penalty function is a nonconcave function to reduce estimation bias and is singular at the origin to reduce model complexity. The nonconcave penalized likelihood approach has several advantages: it deletes an insignificant vari-

able by estimating its coefficient to be zero. Therefore the proposed variable selection procedure actually is an estimation procedure. Compared with the best subset variable selection, the proposed procedure dramatically reduces computational burden and yields a stable model.

In Section 2, we extend the nonconcave penalized approach to analyze data from uniform designs. An example will be illustrated how to apply the extended procedure in practice. We propose a new mathematical formation for analysis of computer experiment in Section 3. The new idea will be demonstrated by a classical example in the literature of analysis of computer experiments. Conclusions are given in Section 4.

2 Analysis of Uniform Designs

Let us start from a typical example of uniform design, analyzed in Fang [10] using a different method and model.

Example 2.1 (**Environmental data**) To study how environmental pollutants affect human health, an experiment was conducted. Environmentalists believe that contents of some metal elements in water would directly affect human health. An experiment using uniform design was conducted in the department of biology, Hong Kong Baptist University. Of interest in this study is the association between the mortality of some kind cell of mice and contents of six metals: Cadmium (Cd), Copper (Cu), Zinc (Zn), Nickel (Ni), Chromium (Cr), and Lead (Pb). To address this issue, the investigator took 17 levels for the content of each metal: 0.01, 0.05, 0.1, 0.2, 0.4, 0.8, 1, 2, 5, 8, 10, 12, 16, 18, 20 (ppm), and then use a uniform design table, listed in Table 1, to construct a 17 times of experiments. Table 1 is a typical table of uniform designs. Like orthogonal designs, such tables are usually tabulated in books by Fang and his coauthors, such as Fang and Wang [13] and Fang [10]. Many uniform design tables can be directly downloaded from world web site at http://www.math.hkbu.edu.edu.hk/UniformDeisgn.

Replacing the symbolic levels in Table 1 by the actual levels, the investigator obtained an actual design in the left panel of Table 2, extracted from Fang [10]. For each combination in the left panel of Table 2, three experiments were conducted, the outputs (mortality) are depicted in the right panel of Table 2, from which it can be seen that the mortality in the last row corresponding high levels of contents of metals is higher than the other ones. This implies that the contents of metals may affect the mortality. After conducting the experiments and collecting the data, the investigator has to analyze the data in order to understand how the mortality associates with the levels of metal contents. ■

Statistical Formulation

Statisticians describe the relationship between the input variables and the output variable by a regression model

$$\text{output variable} = f(\text{input variables}) + \text{random error}, \tag{2.1}$$

where the random error has zero mean and constant variance. In other words, the expectation of outputs is functional associated with the input variables: $f(\text{inputs})$. Very often, it is difficult to

Table 1: Uniform Design $U_{17}(17^6)$

Cd	Cu	Zn	Ni	Cr	Pb
1	4	6	10	14	15
2	8	12	3	11	13
3	12	1	13	8	11
4	16	7	6	5	9
5	3	13	16	2	7
6	7	2	9	16	5
7	11	8	2	13	3
8	15	14	12	10	1
9	2	3	5	7	16
10	6	9	15	4	14
11	10	15	8	1	12
12	14	4	1	15	10
13	1	10	11	12	8
14	5	16	4	9	6
15	9	5	14	6	4
16	13	11	7	3	2
17	17	17	17	17	17

Table 2: Environmental Data

Cd	Cu	Zn	Ni	Cr	Pb	Y_1	Y_2	Y_3
0.01	0.2	0.8	5.0	14.0	16.0	19.95	17.6	18.22
0.05	2.0	10.0	0.1	8.0	12.0	22.09	22.85	22.62
0.1	10.0	0.01	12.0	2.0	8.0	31.74	32.79	32.87
0.2	18.0	1.0	0.8	0.4	4.0	39.37	40.65	37.87
0.4	0.1	12.0	18.0	0.05	1.0	31.90	31.18	33.75
0.8	1.0	0.05	4.0	18.0	0.4	31.14	30.66	31.18
1.0	8.0	2.0	0.05	12.0	0.1	39.81	39.61	40.80
2.0	16.0	14.0	10.0	5.0	0.01	42.48	41.86	43.79
4.0	0.05	0.1	0.4	1.0	18.0	24.97	24.65	25.05
5.0	0.8	4.0	16.0	0.2	14.0	50.29	51.22	50.54
8.0	5.0	16.0	2.0	0.01	10.0	60.71	60.43	59.69
10.0	14.0	0.2	0.01	16.0	5.0	67.01	71.99	67.12
12.0	0.01	5.0	8.0	10.0	2.0	32.77	30.86	33.70
14.0	0.4	18.0	0.2	4.0	0.8	29.94	28.68	30.66
16.0	4.0	0.4	14.0	0.8	0.2	67.87	69.25	67.04
18.0	12.0	8.0	1.0	0.1	0.05	55.56	55.28	56.52
20.0	20.0	20.0	20.0	20.0	20.0	79.57	79.43	78.48

specify or estimate f directly. Hence, statisticians usually approximate f by a linear regression model with a number of unknown regression coefficients. Thus, the task of estimating f becomes to estimate the unknown regression coefficients. As usual, denote the input variables by x's and output variable by y. Model (2.1) was approximated by the following linear regression model

$$y = \beta_1 x_1 + \cdots + \beta_d x_d + \varepsilon,$$

where $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. We can introduce an intercept term by setting $x_1 \equiv 1$. It is worthwhile to note that in order to attenuate modeling bias, we not only include linear terms of input variables, but also many related variables, such as quadratic terms and interaction terms. Thus, the x_i can be the linear terms of input variables as well as the quadratic terms, interaction terms of input variables. Therefore, the number of regression coefficients, d, may be much larger than the number of input variables. Very frequently, many x variables in the model are not significant, and therefore, should be excluded from the model in order to obtain a simple and easily interpreted model.

Variable Selection via Penalized Least Squares

For ease of presentation, matrix notation will be used. Let y_i be the output of the experiment whose associated x variable is $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^T$. Denote $\mathbf{y} = (y_1, \cdots, y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$. Many variable selection criteria for the linear regression model are closely related with the penalized least squares

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|), \tag{2.2}$$

where $\|\mathbf{a}\| = (\mathbf{a}^T\mathbf{a})^{1/2}$, $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a tuning parameter, controlling model complexity and chosen by a data-driven method, such as cross-validation (CV) and generalized cross validation (GCV, Craven and Wahba [6]).

Take the penalty function to be the entropy penalty, namely, $p_{\lambda_n}(|\theta|) = \frac{1}{2}\lambda_n^2 I(|\theta| \neq 0)$, which is also referred to as L_0-penalty in the literature, where $I(\cdot)$ is an indicator function. Note that the dimension or the size of a model equals to the number of nonzero regression coefficients in the model. Hence, many popular variable selection criteria can be derived from the penalized least squares (2.2) by setting different values for λ_n. For instance, the AIC (Akaike [1], or C_p Mallows [20]), BIC, ϕ-criterion (Hannan and Quinn [16] and Shibata [25]) and RIC (Foster and George [14]) correspond to $\lambda_n = \sqrt{2}(\sigma/\sqrt{n})$, $\sqrt{\log n}(\sigma/\sqrt{n})$, $\sqrt{\log\log n}(\sigma/\sqrt{n})$ and $\sqrt{2\log(d)}(\sigma/\sqrt{n})$, respectively, although these criteria were motivated from different principles. Since the entropy penalty function is discontinuous, it requires searching over all possible subsets for finding the solution of this penalized least squares. This is very expensive in terms of computational cost. Furthermore, the resulting model is unstable (see Breiman [4]).

The family of L_p penalty $p_{\lambda_n}(|\theta|) = \lambda_n p^{-1}|\theta|^p$ has been used for the penalized least squares. The L_2 penalty results in a ridge regression estimator. The L_p $(0 < p < 1)$ penalty yields the bridge regression (Frank and Friedman [15]). The non-negative garrote (Breiman [5]) is closely related

to the penalized least squares with the L_p penalty ($p < 1$). The L_1 penalty yields the LASSO (Tibshirani [26]),

Advocated by Fan and Li [8], to achieve the purpose of variable selection for linear models, a good penalty function should yield an estimator with three properties: (a) **sparsity**: the resulting estimator may automatically set small estimated coefficients to be zero in order to reduce model complexity. In other words, the resulting estimator should be a thresholding rule. (b) **unbiasedness**: the resulting estimator is nearly unbiased when the true unknown coefficient is large in order to avoid unnecessary modeling bias; (c) **continuity**: the resulting estimator should be continuous in some sense in order to avoid instability in model prediction.

Unfortunately, none of the aforementioned penalty functions satisfy the mathematical requirements for the three properties. Fan and Li [8] suggest to use the smoothly clipped absolute deviation (SCAD) penalty, whose derivative is given by

$$p'_\lambda(\theta) = \lambda I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{a-1} I(\theta > \lambda) \text{ for some } a > 2 \quad \text{and} \quad \theta > 0 \tag{2.3}$$

and $p_\lambda(0) = 0$, satisfies all the three mathematical conditions. We use the acronym SCAD for all procedures using the SCAD penalty. The SCAD involves two unknown parameters λ and a, where λ is a regularization parameter chosen by the generalized cross validation (GCV), a data-driven method. Fan and Li [8] suggested using $a = 3.7$ based on a Bayesian argument. Hence, this value will be used throughout the whole paper.

Oracle Property

Assume that observations are taken from the following model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Without loss of generality, assume that all components of the first portion ($\mathbf{X}_1$) are active, while the second portion ($\mathbf{X}_2$) is not active. An ideal estimator is the oracle estimator:

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}, \qquad \text{and} \qquad \widehat{\boldsymbol{\beta}}_2 = \mathbf{0}.$$

The oracle estimator correctly specifies the true model and efficiently estimate the regression coefficient of $\mathbf{X}_1$. This is a desired property in variable selection. With a proper rate of λ_n, Li and Lin [19] showed the SCAD possesses the oracle property in asymptotic sense. Thus, from theoretic point of view, the SCAD is an ideal variable selector.

An Algorithm

It is challenging in finding solutions of nonconvex penalized least squares, because the target function is nonconvex and could be high-dimensional. Furthermore, the penalized least squares function with the SCAD do not have the second derivative at some points due to the penalty function. In order to apply the Newton Raphson algorithm to the penalized least squares, we locally approximate the SCAD by a quadratic function as follows.

Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the true value of $\boldsymbol{\beta}$, when $\beta_j^{(0)}$ is not very close to 0, the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by the quadratic function as

$$[p_{\lambda_n}(|\beta_j|)]' = p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \approx \{p'_{\lambda_n}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j, \tag{2.4}$$

otherwise, set $\widehat{\beta}_j = 0$. With the local quadratic approximation, the solution for the penalized least squares can be found by iteratively computing the following ridge regression with an initial value $\boldsymbol{\beta}^{(0)}$:

$$\boldsymbol{\beta}^{(1)} = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}^{(0)})\}^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.5}$$

where

$$\Sigma_\lambda(\boldsymbol{\beta}^{(0)}) = \text{diag}\{p'_{\lambda_1}(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \cdots, p'_{\lambda_d}(|\beta_d^{(0)}|)/|\beta_d^{(0)}|\}.$$

Choice of tuning parameter λ_n

From the iterative ridge regression (2.5), the fitted value of $\mathbf{y}$ is

$$\widehat{\mathbf{y}} = \mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_{\lambda_n}(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Although $\widehat{\mathbf{y}}$ is not a linear in term of $\mathbf{y}$,

$$\mathbf{P}_\mathbf{X}\{\widehat{\boldsymbol{\beta}}(\lambda)\} = \mathbf{X}\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\widehat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}^T$$

can be regarded as a projection matrix. Thus, a given λ,

$$e(\lambda) = \text{tr}[\mathbf{P}_\mathbf{X}\{\widehat{\boldsymbol{\beta}}(\lambda)\}]$$

can be regarded as effective number of parameters. The generalized cross-validation statistics can be defined as by

$$\text{GCV}(\lambda) = \frac{\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\|^2}{\{n - e(\lambda)\}^2} \tag{2.6}$$

and we estimate λ_n by $\widehat{\lambda}_n = \text{argmin}_\lambda\{\text{GCV}_a(\lambda)\}$.

A Variable Selection Procedure

For a given value of λ_n and an initial value of $\boldsymbol{\beta}$, we iteratively compute the ridge regression (2.5) with updating local quadratic approximation (2.4) at each step during iteration. This can be easily implemented in many statistical packages. Some components of the resulting estimate will exactly be 0. These components correspond to the coefficient of inactive effects. In other words, nonzero components of the resulting estimate are correspondent to the active effects in the SSD.

Compared with stepwise regression, the proposed procedure is actually an estimation procedure rather than a variable selection procedure. Compared with the best subset variable selection, our procedure significantly reduces the computational burden. Moreover, it simultaneously excludes inactive factors and estimates the regression coefficient of active effects.

Analysis of Example 2.1

Note that the ratio of maximum to minimum of contents of six metals is 2000. To stabilize numerical computation, all x variables were standardized first. Denote $x_1, ..., x_6$ to the standardized variables of Cd, Cu, Zn, Ni, Cr and Pb, respectively. That is, for example, $x_1 = (\text{Cd} - 6.5624)/7.0656$. All linear terms, quadratic terms and interactions between the linear terms are chosen as x variables in the linear model. Thus, including an intercept term, there are 28 predictors in the model. The model is overfitted because totally there are only 17 design combinations in this experiment. Thus, variable selection is necessary. The variable selection procedure is applied for this data set. the estimated λ equals 0.075. The estimated coefficients of x variables (rather than the original scale of contents) are depicted in Table 3. Total 12 variables are included in the final model. Note that $t_{0.005}(38) = 2.7116$. All selected variable are very statistically significant. From Table 3, it can be seen that the effect of Cd is quadratic, the positive effect of Cu and Ni is linear, in addition, Cu & Cr and Ni & Cr have negative interaction. Cr has negative effect, further, Cr has negative interaction with Cu, Zn and Ni. The effect of Zn is quadratic. Finally, Pb has positive linear and quadratic effect. Moreover, Pb and Zn has a positive interaction.

Table 3: Estimates, Standard Errors and t-value

X-variable	Estimate	Standard Error	$\lvert t\rvert$
Intercept	36.4539	0.5841	62.4086
Cd	14.9491	0.2944	50.7713
Cu	12.8761	0.2411	53.4060
Ni	0.9776	0.2510	3.8950
Cr	-7.2696	0.2474	29.3900
Pb	4.0646	0.2832	14.3536
Cd^2	-6.2869	0.3624	17.3480
Zn^2	2.8666	0.3274	8.7554
Pb^2	9.2251	0.4158	22.1856
Cu*Cr	-1.6788	0.3171	5.2945
Zn*Cr	-6.2955	0.3306	19.0401
Zn*Pb	11.9110	0.2672	44.5708
Ni*Cr	-11.3896	0.4303	26.4680

3 Analysis of Computer Experiments

Notice that the outputs of computer experiments are deterministic, therefore we describe the relationship between the input variables and the output variable by the following model

$$\text{output variable} = f(\text{input variables}). \tag{3.1}$$

Compared with model (2.1), model (3.1) does not have a random error term. Hence, the techniques related with analysis of linear regression cannot be applied directly for model (3.1). Therefore, it poses many challenges for statisticians to analyze data from design of computer experiments.

Linear regression and Bayesian kriging are the two main approaches to modeling f in statistical literature. Koehler and Own [18] gave an excellent review on this topic. Linear regression approximates f by a linear combination of a set of bases of the functionals space of f. An and Owen [2], Fang and Lin [12] and Jiang and Owen [17] illustrate how to implement linear regression techniques to find a good approximation of f. A Bayesian approach to modeling f is based on spatial model adapted from the geo-statistical kriging model (Matheron [21]), also see Cressie [7]. The Bayesian approach regards the bias, or systematic departure of the response surface from a linear model, as the realization of a Gaussian process. The advantage of this approach is that the resulting model has exact predictions at the observed responses and predicts the unobserved responses by a linear combination of the observations. As is common with Bayesian methods, there may be difficult in finding an appropriate prior distribution. To implement the Bayesian kriging approach, one has to specify the covariance function of the underlying Gaussian process. Sacks, Welch, Mitchell and Wynn [24] discuss in details how to implement the Bayesian kriging approach for analysis of computer experiment.

In this section, we present a new mathematical formulation for analysis of computer experiments. Our formulation provides insights into why the linear regression approach may perform well in the analysis of computer experiments.

3.1 Regularization interpolation and penalized least squares

Modeling f indeed is an interpolation problem in mathematics. Of course, there are many ways to interpolate the outputs. Let $b_1(\mathbf{x}), b_2(\mathbf{x}), \cdots$, be a set of bases of the functional space $\mathcal{F}$ of f, and $\mathbf{b}_i = (b_i(\mathbf{x}_1), \cdots, b_i(\mathbf{x}_n))^T$ for $i = 1, 2, \cdots$. Let B be $(\mathbf{b}_1, \cdots, \mathbf{b}_N)$. Denote $\mathbf{f_n} = (\mathbf{f(x_1)}, \cdots, \mathbf{f(x_n)})^{\mathbf{T}}$ be observed outputs at design points $\mathbf{x}_1, \cdots, \mathbf{x}_n$. To interpolate f using the bases $b_1(\mathbf{x}), b_2(\mathbf{x}), \cdots$, we take N is large enough such that the following equation has a solution:

$$\mathbf{f_n} = \mathbf{B}\boldsymbol{\gamma}. \tag{3.2}$$

It is worthwhile to noting that the bases for the functional space $\mathcal{F}$ can be constructed via many different ways, among which Fourier bases, wavelets bases and spline bases are three popular approaches. In order to obtain a better interpolation to f, we usually take N is quite large. This leads to the equation (3.2) become an underdetermined system of equations, there are many different solutions for $\boldsymbol{\gamma}$ that match sample data $\mathbf{f_n}$. To deal with this issue, statisticians choose a $\boldsymbol{\gamma}$ that minimizes $\|\boldsymbol{\gamma}\|^2$ under the constraint (3.2). This method was referred as the *normalized method of frame* in statistical literature (Rao [23] and Antoniadis and Fan [3]).

Thus, the interpolation problem is equivalent to minimizing

$$\boldsymbol{\gamma}^T\boldsymbol{\gamma} + \lambda_0\|B\boldsymbol{\gamma} - \mathbf{f_n}\|^{\mathbf{2}},$$

where λ_0 is a Lagrange multiplier. Let $\mathbf{y} = (y_1, \cdots, y_n)$ be the output vector of computer experiments. Thus, $\mathbf{y} = \mathbf{f_n}$ as the outputs of computer experiment are deterministic. The minimization problem is equivalent to the penalized least squares problem

$$\frac{1}{2}\|\mathbf{y} - B\boldsymbol{\gamma}\|^2 + \lambda\boldsymbol{\gamma}^T\boldsymbol{\gamma} \tag{3.3}$$

which is a special case of the penalized least squares (2.2).

The goal of computer experimental designs is to build a model for f with a good prediction over the experimental domain and to find an optimal design point over the whole experimental domain. It is desirable to have a simple form of the estimated f for the purpose of optimization and prediction. This motivates us to pursue a less complicated model than the resulting model in (3.3). To obtain a less complicated model, we allow the resulting estimate to have a small departure from the output, i.e., $\|\mathbf{y} - B\boldsymbol{\gamma}\| \leq \varepsilon$, rather than exact interpolation, $\|\mathbf{y} - B\boldsymbol{\gamma}\| = 0$. Thus, we may employ the variable selection techniques in linear regression model to find a less complicated model. Motivated from this point of view, we apply the ideas of penalized least squares introduced in Section 2 for the analysis of computer experiments. That is, we find a solution of $\boldsymbol{\gamma}$ by minimizing

$$\frac{1}{2}\|\mathbf{y} - B\boldsymbol{\gamma}\|^2 + n\sum_{j=1}^{N} p_\lambda(|\gamma_j|). \tag{3.4}$$

We illustrate this extension via an application of a classical example in the field of computer experiments.

3.2 An illustration

In this section, we revisit a classical example of computer experiments. This example has been studied by many authors, for instance, Worley [28], Morris, Mitchell and Ylvisaker [22], An and Owen [2], and Fang and Lin [12]. See the references for a detailed description.

Example 3.1 (**Borehole function**) The response variable y, is determine by

$$y = \frac{2\pi T_u[H_u - H_l]}{\ln(\frac{r}{r_w})\left[1 + \frac{2LT_u}{\ln(r/r_w)r_w^2 K_w} + \frac{T_u}{T_l}\right]},$$

where the 8 input variables have an experimental domain:

$r_w \in [0.05\ ,\ 0.15]$, $r \in [100\ ,\ 50000]$, $T_u \in [63070\ ,\ 115600]$, $T_l \in [63.1\ ,\ 116]$,

$H_u \in [990\ ,\ 1110]$, $H_l \in [700\ ,\ 820]$, $L \in [1120\ ,\ 1680]$, $K_w \in [9855\ ,\ 12045]$.

A $U_{30}(30^8)$, listed in Table 4 and downloaded from the website of uniform design, was employed to generate a design with 30 experiments, and the outcomes and their corresponding design points are depicted in Table 5. As is listed in Table 5, denote x_1 to be r_w, so on. The means and standard deviations of x_i's are displayed in Table 6. Let z_i be the standardized variable of x_i. We construct quadratic splines $(z_i - k_j)_+^2$ with the knots at the median of z_i, denoted by k_2, and lower and upper 20 percentiles of z_i, denoted by k_1 and k_3. The knots of z_i's are also depicted in Table 6. Here a_+ mean the positive part of a and is taken before taking square in the construction of the quadratic spline basis. Motivated from Fang and Lin [12], we take natural logarithm of outputs as response variable y in model (3.3). For each variable z_i, $i = 1, \cdots, 8$,

$$z_i, z_i^2, (z_i - k_1)_+^2, (z_i - k_2)_+^2, (z_i - k_3)_+^2,$$

are introduced to the initial model. Note that k_j, $j = 1, 2, 3$ may be different from one x variable to another x variable, and their values are depicted in Table 6. We also include an intercept and all

Table 4: Uniform Design $U_{30}(30^8)$

No.	r_w	r	T_u	T_l	H_u	H_l	L	K_w
1	4	28	13	25	30	16	12	12
2	24	22	18	11	26	23	7	2
3	14	7	11	28	7	24	6	6
4	12	14	30	3	5	20	8	19
5	3	9	19	5	20	28	24	13
6	6	3	15	15	23	17	1	27
7	10	17	25	27	25	26	17	22
8	26	12	8	4	11	10	2	11
9	7	25	24	7	3	12	14	5
10	17	24	6	22	19	29	3	18
11	8	13	21	21	12	14	30	1
12	27	11	10	18	28	21	29	24
13	21	20	7	26	22	6	23	4
14	28	27	12	6	18	2	18	20
15	18	8	29	14	29	8	20	7
16	19	4	4	1	17	19	16	3
17	16	2	14	12	1	4	28	15
18	20	30	20	16	8	7	5	23
19	9	15	1	8	27	5	9	17
20	22	5	26	9	13	27	13	25
21	11	29	3	13	14	25	27	10
22	15	23	22	2	24	11	26	28
23	5	6	5	23	6	9	22	21
24	2	19	27	19	16	3	4	9
25	1	21	9	10	9	22	19	26
26	29	1	23	24	21	13	10	14
27	13	10	17	30	15	1	15	30
28	23	18	2	20	2	15	11	29
29	30	16	16	17	4	30	21	8
30	25	26	28	29	10	18	25	16

interactions of z_i's in the initial model. Thus, there are totally 68 dependent variables in the initial model. Since we only conduct 30 experiments, the initial model is over-parameterized. The variable selection strategy introduced in Section 2 is applied for this example. Detailed implementation procedure of variable selection is given in Li and Lin [12]. The estimated λ is $1.1923 * 10^{-5}$. The solution of penalized least squares (3.3) is depicted in Table 7.

As usual, *Root Mean Square Error* (RMSE) is used to assess the accuracy of the approximation. The RMSE is defined by

$$\text{RMSE} = [E\{f(\mathbf{x}) - B(\mathbf{x})\widehat{\boldsymbol{\gamma}}\}^2]^{1/2},$$

Table 5: Designs and outputs

$x_1 = r_w$	$x_2 = r$	$x_3 = T_u$	$x_4 = T_l$	$x_4 = H_u$	$x_6 = H_l$	$x_7 = L$	$x_8 = K_w$	output
0.0617	45842	84957	106.3017	1108	762	1335	9990	30.8841
0.1283	35862	93712	81.6150	1092	790	1241	10136	126.2840
0.0950	10912	81456	111.5917	1016	794	1223	10063	51.6046
0.0883	22555	114725	67.5083	1008	778	1260	10048	44.7063
0.0583	14238	95464	71.0350	1068	810	1559	9983	17.6309
0.0683	4258	88460	88.6683	1080	766	1129	10005	40.7011
0.0817	27545	105970	109.8283	1088	802	1428	10034	41.9919
0.1350	19228	76203	69.2717	1032	738	1148	10151	146.8108
0.0717	40852	104218	74.5617	1000	746	1372	10012	29.8083
0.1050	39188	72701	101.0117	1064	814	1167	10085	74.3997
0.0750	20892	98965	99.2483	1036	754	1671	10019	29.8223
0.1383	17565	79704	93.9583	1100	782	1652	10158	116.6914
0.1183	32535	74451	108.0650	1076	722	1540	10114	101.7336
0.1417	44178	83207	72.7983	1060	706	1447	10165	154.9332
0.1083	12575	112974	86.9050	1104	730	1484	10092	93.2778
0.1117	5922	69199	63.9817	1056	774	1409	10100	78.5678
0.1017	2595	86708	83.3783	992	714	1633	10078	55.4821
0.1150	49168	97215	90.4317	1020	726	1204	10107	101.7270
0.0783	24218	63945	76.3250	1096	718	1279	10027	56.9115
0.1217	7585	107721	78.0883	1040	806	1353	10121	80.7530
0.0850	47505	67447	85.1417	1044	798	1615	10041	34.6025
0.0983	37525	100717	65.7450	1084	742	1596	10070	65.1636
0.0650	9248	70949	102.7750	1012	734	1521	9997	24.2095
0.0550	30872	109472	95.7217	1052	710	1185	9975	27.3042
0.0517	34198	77954	79.8517	1024	786	1465	9968	13.5570
0.1450	932	102468	104.5383	1072	750	1297	10173	165.6246
0.0917	15902	91961	115.1183	1048	702	1391	10056	65.8352
0.1250	29208	65697	97.4850	996	758	1316	10129	89.2366
0.1483	25882	90211	92.1950	1004	818	1503	10180	86.2577
0.1317	42515	111222	113.3550	1028	770	1577	10143	89.7999

where the expectation is taken with respect to $\mathbf{x}$ assuming to have the uniform distribution over the design domain. The RMSE can be estimated by

$$\widehat{\mathrm{RMSE}} = \left[\frac{1}{N}\sum_{l=1}^{N}\{f(\mathbf{x}_l) - B(\mathbf{x}_l)\widehat{\boldsymbol{\gamma}}\}^2\right]^{1/2}$$

by taking N large enough and $\mathbf{x}_l$'s is a random sample from the uniform distribution over the design domain. In this example, we take $N = 10000$, the estimated RMSE of the resulting model is 0.3335.

Table 6: means, standard deviations and knots

Variable	Mean	Std	k_1	k_2	k_3
x_1	-2.3478	0.3129	-1.0421	0.1439	1.0072
x_2	9.8398	0.9441	-0.7144	0.3054	0.8152
x_3	89335	15415	-1.0564	0	1.0564
x_4	89.5500	15.5233	-1.0564	0	1.0564
x_5	1050.0	35.2136	-1.0564	0	1.0564
x_6	760.0	35.2136	-1.0564	0	1.0564
x_7	1400.0	164.3	-1.0564	0	1.0564
x_8	10074	64.2648	-1.0564	0	1.0564

Table 7: Estimate of γ_i's

Variable	Estimated Coefficients
intercept	4.0864
z_1	0.6289
z_3	-0.0041
z_4	0.0010
z_5	0.1231
z_6	-0.1238
z_7	-0.1179
z_5^2	-0.0095
z_6^2	-0.0072
z_7^2	0.0067
$z_1 z_2$	0.0005
$z_1 z_4$	0.0024
$z_2 z_3$	-0.0005
$z_4 z_6$	0.0007
$z_5 z_6$	0.0165
$z_5 z_7$	0.000035
$z_6 z_7$	-0.0011
$(z_2 - k_1)_+^2$	-0.0007
$(z_3 - k_1)_+^2$	0.0021
$(z_4 - k_1)^2$	-0.0004
$(z_2 - k_2)_+^2$	-0.0002
$(z_5 - k_2)_+^2$	-0.0127
$(z_7 - k_2)_+^2$	0.0014
$(z_8 - k_2)_+^2$	0.0275
$(z_1 - k_3)_+^2$	-0.0174
$(z_2 - k_3)_+^2$	-0.0014

Worley [28] conducted a design with 10 experiments for this borehole example by using three different approaches. He estimated f using both observed values of y and its derivatives. The estimated RMSEs for his resulting three models are 1.89, 2.45 and 2.37 over 50 test sites, respectively. Morris, Mitchell and Ylvisaker [22] constructed a design with 10 experiments for this borehole function. They used both observed values and its first derivatives. They concluded a model with an estimated MSE 0.61 over 50 test sites, Fang, Ho and Xu [11] constructed a uniform design with 32 experiments, they use a quadratic regression model to approximate the borehole function. Their model has an estimated RMSE 0.5077. With the same uniform design, the B-spline model used in Fang and Lin [12] has an estimated RMSE 2.1095. Thus, compared with the results in the literature, the quadratic spline model used in this paper outperforms the others. It is worth to noting that we used a uniform design with 30 experiments. This is because the uniform design is effective and robust to the model used for approximation.∎

4 Conclusions

In this paper, we introduced a variable selection procedure via nonconvex penalized least squares approach for analysis of uniform design and analysis of computer experiment. The procedure was illustrated via two applications of uniform design and design of computer experiment. we present a new mathematical formulation for analysis of computer experiment. This formulation provides insight into why the linear regression with various variable selection procedures can be applied for the analysis of computer experiments.

References

1. H. Akaike (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, **19**, 716–723.
2. J. An and A. B. Owen (2001). Quasi-regression. *J. Complexity.* **17**. In press.
3. A. Antoniadis and J. Fan (2001). Regularization of wavelets approximations (with discussions). *Journal of American Statistical Association*, **96**, 939-967.
4. L. Breiman (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373-384.
5. L. Breiman (1996). Heuristics of instability and stabilization in model selection, *Ann. Statist.*, **24**, 2350-2383.
6. P. Craven and G. Wahba (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.
7. N. A. Cressie (1993). *Statistics for Spatial Data (Revised edition)*, Wiley, New York.
8. J. Fan and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association.* **96**, 1348-1360.
9. K. T. Fang (1980). The uniform design: application of number-theoretic methods in experimental design. *Acta Math. Appl. Sinica*, **3**, 363-372.
10. K. T. Fang (1994). *Uniform Design and Uniform Design Tables.* Science Press, Beijing.

11. K. T. Fang, W. M. Ho and Z. Q. Xu (2000). Case studies of computer experiments with uniform design. Manuscript.
12. K. T. Fang and D. K. J. Lin (2001). Uniform experimental design and its application in industry. *Handbook of Statistics*, eds C. R. Rao.
13. K. T. Fang and Y. Wang (1994). *Applications of Number Theoretic Methods in Statistics*, Chapman and Hall, London.
14. D. P. Foster and E. I. George (1994). The risk inflation criterion for multiple regression, *Annals of Statistics.*, **22**, 1947-1975.
15. I. E. Frank and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.
16. E. J. Hannan and B. G. Quinn (1979). The determination of the order of autoregression, *Journal of the Royal Statistical Society, Ser. B*, **41**, 190-195.
17. T. Jiang and A. B. Owen (2001). Quasi-regression with shrinkage. Manuscript.
18. J. R. Koehler and A. B. Owen (1996). Computer experiments, in *Handbook of Statistics*, Vol. 13, Eds by S. Ghosh and C. R. Rao, Elsevier Science B. V., Amsterdam, 261-308.
19. R. Li and D. K. J. Lin (2002). Analysis of Supersatured Design. *Statistics & Probability Letters*. To appear.
20. C. L. Mallows (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.
21. G. Matheron (1963). Principles of geostatistics, *Econm. Geol.*, **58**, 1246-1266.
22. D. M. Morris, T. J. Mitchell and D. Ylvisaker (1993). Bayesian design and analysis of computer experimental: use of derivatives in surface prediction, *Technometrics*, **35**, 243-255.
23. C. R. Rao (1973). *Linear Statistical Inference and its Applications*, Wiley, New York.
24. J. Sacks, W. J. Welch, T. J. Mitchell and H. P. Wynn (1989). Design and analysis of computer experiments. *Statist. Sci.*, **4**, 409-435.
25. R. Shibata (1984). Approximation efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43-49.
26. R. J. Tibshirani (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistical Society, B*, **58**, 267-288.
27. Y. Wang and K. T. Fang (1981). A note on uniform distribution and experiment design. *KeXue TongBao*, **26**, 485-489.
28. B. A. Worley (1987). Deterministic uncertainty analysis, ORN-0628, available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161, USA.

About the Authors

R. Li is an assistant professor at Department of Statistics, The Pennsylvania State University at University Park.