Submitted to Technometrics

ANALYSIS OF COMPUTER EXPERIMENTS USING PENALIZED LIKELIHOOD GAUSSIAN KRIGING MODEL

Runze Li
Department of Statistics
Pennsylvania State University

Agus Sudjianto*
V-Engine Engineering Analytical Powertrain
Ford Motor Company

* Corresponding Author
POEE Bld., FK143 (Pillar M13), MD 65
21500 Oakwood Blvd., Dearborn, MI 48121-4091
E-mail: asudjian@ford.com

ABSTRACT

Kriging is a popular analysis approach for computer experiment for the purpose of creating a cheap-to-compute "metamodel" as a surrogate to a computationally expensive engineering simulation model. The maximum likelihood approach is employed to estimate the parameters in the Kriging model. However, the likelihood function near the optimum may be flat in some situations, and this leads to the maximum likelihood estimate for the parameters in the covariance matrix to have a very large random variation. To overcome this difficulty, a penalized likelihood approach is proposed for the kriging model. The proposed method is particularly important in the context of a computationally intensive simulation model where the number of simulation runs must be kept small. We applied the proposed approach for the reduction of piston slap, an unwanted engine noise due to piston secondary motion. Issues related to practical implementation of the proposed approach are discussed.

Keywords: Computer experiment, kriging, penalized likelihood, Fisher Scoring algorithm, metamodel, Smoothly Clipped Absolute Deviation.

1. INTRODUCTION

This research was motivated by ever decreasing product development time where decisions must be made quickly in the early design phase. While sophisticated engineering computer simulations become ubiquitous tools to investigate complicated physical phenomena, their effectiveness to support timely design decisions in quick pace product development is often hindered due to their excessive requirements for model preparation, computational, and output post-processing. The computational requirement increases dramatically when the simulation models are used for the purpose of probabilistic design optimization for which a "double loop" procedure is usually required (Wu and Wang, 1998; Du and Chen, 2002; Kalagnanam and Diwekar, 1997) as illustrated in Figure 1.

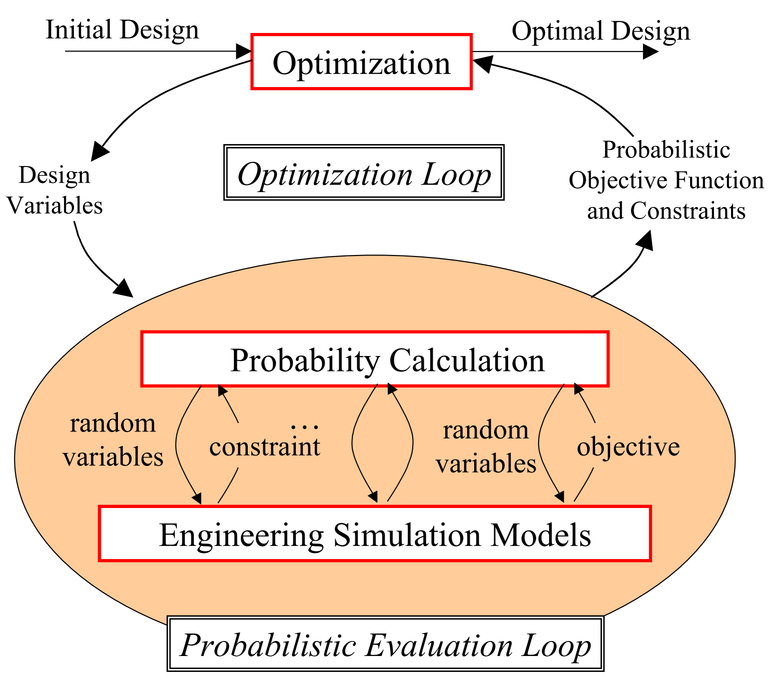

Figure 1. Double loop procedure in probabilistic design

The outer loop is the optimization itself and the inner loops are probability calculations for design objective and design constraint. The most challenging issue for implementing probabilistic design is associated with its *intensive computational demand* of this double loop procedure. To deal with this issue, metamodeling—"model of the model" (Kleijnen, 1987)—to replace an expensive simulation approach becomes a popular choice in many engineering applications (e.g., Booker et al., 1999 ; Hoffman et al., 2003; Du et al., 2003).

The accuracy of metamodels to represent the original model is influenced by both the experimental designs used (see for example, Ye et al., 2000) as well as the metamodeling approach itself (Jin et al., 2000). The topic of design and analysis of computer experiment for metamodeling has recently received a lot of interests from both engineering and statistical communities (Welch et al., 1992; Kohler and Owen, 1996; Jin et al., 2000; Simpson et al., 2002). Since the output obtained from a computer experiment is deterministic, it imposes a challenge in analyzing such data. Many complex methods to analyze outputs of computer models have been proposed in the statistical literature. Review papers (Kohler and Owen, 1996; Sack et al., 1990) provide a detailed review on how to scatter computer design points over the experimental domain effectively and how to analyze the deterministic output. Sack et al. (1990) advocated to model the deterministic output as a realization of a stochastic process, and employed Gaussian stochastic kriging method to predict the deterministic outputs. In implementation of Gaussian kriging models, one may introduce some parameters in the covariance matrix, and the maximum likelihood approach can be used to construct an estimate for the parameters.

Although the Gaussian kriging method is useful and popular in practice (e.g., Booker et al., 1999; Jin et al., 2000; Kodilayam et al., 2001; Meckesheimer et al., 2002; Simpson et al., 2002), it does have some limitations. From our experience, one of the serious problems with the

Gaussian kriging models is that the maximum likelihood estimate for the parameters in the covariance matrix may have very large random variation because the likelihood function near the optimum is flat. We demonstrate this problem in the following example.

Consider the following one-dimensional function

$$y = \sin(x) \tag{1}$$

Let the sample data be $x = \{0, 2, 4, \ldots, 10\}$. Let us agree to use the following Gaussian kriging model to fit the data:

$$y(x) = \mu + z(x) \tag{2}$$

where $z(x)$ is a Gaussian process with mean zero and covariance

$$\mathrm{cov}\{z(s), z(t)\} = \sigma^2 \exp\{-\theta |s - t|^2\} \tag{3}$$

For a given θ, the maximum likelihood estimate for μ and σ^2 can be easily computed. We can further compute the *profile* likelihood function $l(\theta)$, which equals to the maximum of the likelihood function over μ and σ^2 for any given θ. The corresponding logarithm of profile likelihood (log-likelihood, for short) function $l(\theta)$ versus θ is depicted in Figure 2(a) from which we can see that the likelihood function becomes almost flat for $\theta \geq 1$. The prediction based on the Gaussian kriging model is displayed in Figure 2(b), which shows that the prediction becomes very erratic when x is not equal to the sample data.

To avoid the erratic behavior, we consider a penalized likelihood approach, which will be described in detail in Section 2. A penalized log-likelihood function with the SCAD penalty (see Section 2 for its definition) is depicted in Figure 2(c), and its corresponding prediction is displayed in Figure 2(d). Figure 2(c) clearly shows that the penalized likelihood function is not flat around the optimum. The shape of the penalized log-likelihood function implies that the resulting estimate for θ possesses smaller standard error. From Figure 2(d), the prediction and the true curve is almost identical.

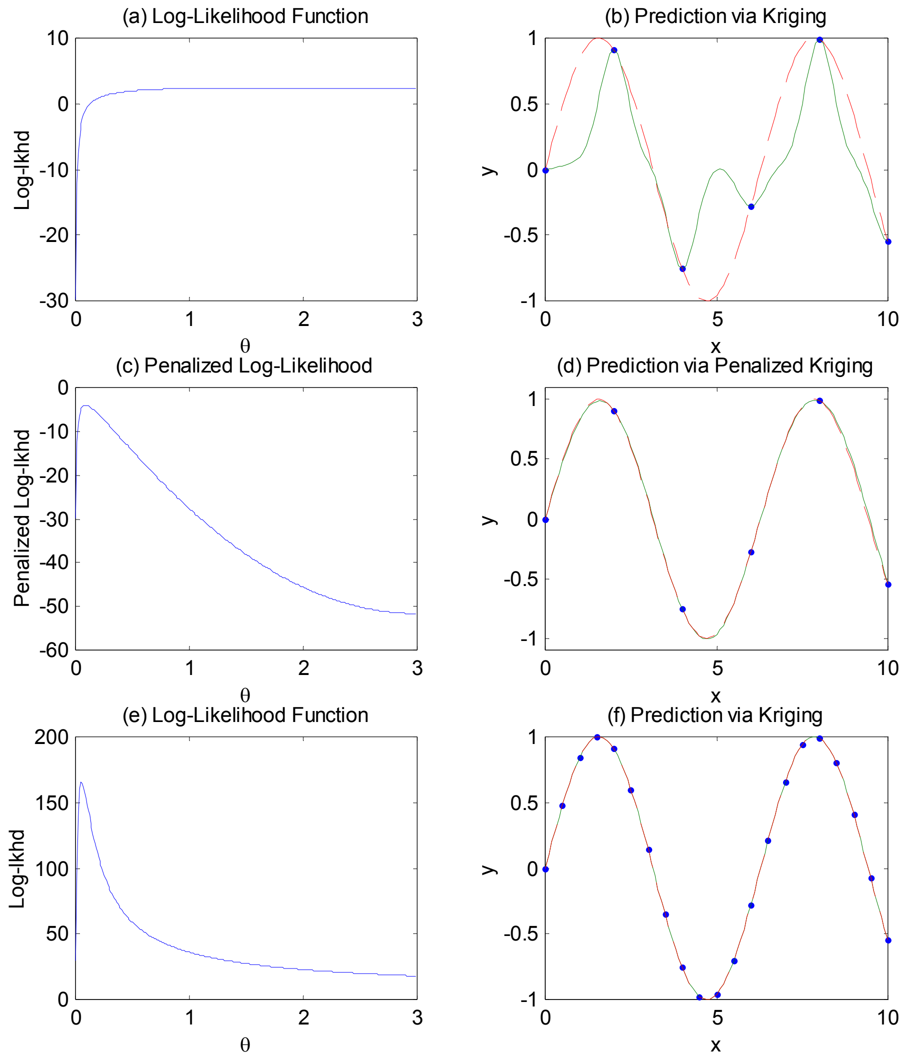

Figure 2. Kriging and Penalized Kriging. (a) and (e) are log-likelihood functions of kriging with sample size $N = 6$ and 21, respectively. (b) and (f) are prediction via kriging when sample size $N = 6$ and 21, respectively. (c) is penalized log-likelihood function of kriging with $N = 6$. (d) is the prediction via penalized kriging with $N = 6$. In (b), (d) and (f), solid line stands for prediction, dashed line stands for true curve, dot stands for prediction at the sample datum points

To demonstrate the effect of sample data to the likelihood function and to properly predict at unsampled point, we consider a slightly larger sample $x = \{0, 0.5, 1, \ldots, 10\}$. The corresponding likelihood function is depicted in Figure 2(e), and the prediction is shown in Figure 2(f). Compare Figure 2(c) with Figure 2(e), the locations of the maximum likelihood estimate and the penalized maximum likelihood estimate are very close. Furthermore, Figure 2(f) confirms that the corresponding prediction yielded by the penalized kriging method is a good prediction.

In this paper, a new approach via penalized likelihood is proposed to model the outputs of computer experiments. We further discuss the choice of penalty functions. Using a simple approximation to the penalty function, the proposed method can be easily carried out with *Fisher* scoring algorithm. Furthermore, we proposed a method to choose the regularization parameter involved in the penalty function. We summarize the proposed approach as an easy-to-follow algorithm. The benefit of the proposed method is demonstrated using an engineering example of the design of power conversion to minimize piston slap noise.

2. PENALIZED LIKELIHOOD GAUSSIAN KRIGING MODELS

Suppose that $\mathbf{x}_i$, $i = 1, \ldots, N$, are design points over a d-dimensional experimental domain $\mathcal{D}$, and $y_i = y(\mathbf{x}_i)$ are sampled from the model

$$y(\mathbf{x}_i) = \mu + z(\mathbf{x}_i) \tag{4}$$

where $z(\mathbf{x}_i)$ is a Gaussian process with zero mean and covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$r\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma^2 \exp\left\{-\sum_{k=1}^{d} \theta_k \left|x_{ik} - x_{jk}\right|^q\right\}, \quad 0 < q \le 2, \tag{5}$$

where $\theta_k \geq 0$. Let $\boldsymbol{\gamma} = (\theta_1, \ldots, \theta_d, \sigma^2)^{\mathrm{T}}$ and define $\mathbf{R}(\boldsymbol{\gamma})$ to be an $N\times N$ matrix with the (i, j)-element $r(\mathbf{x}_i, \mathbf{x}_j)$. Thus, the density of $\mathbf{y} = (y_1, \ldots, y_N)^{\mathrm{T}}$ is

$$f(\mathbf{y}) = (2\pi)^{-N/2} \left|\mathbf{R}(\boldsymbol{\gamma})\right|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{1}_N\,\mu)^{\mathrm{T}}\,\mathbf{R}^{-1}(\boldsymbol{\gamma})(\mathbf{y}-\mathbf{1}_N\,\mu)\right\} \tag{6}$$

where $\mathbf{1}_N$ is an N-dimensional vector with all elements equaling 1. After dropping a constant, the log-likelihood function of the collected data equals

$$l(\mu,\boldsymbol{\gamma}) = -\frac{1}{2}\log\left|\mathbf{R}(\boldsymbol{\gamma})\right|^{-1/2} - \frac{1}{2}(\mathbf{y}-\mathbf{1}_N\,\mu)^{\mathrm{T}}\,\mathbf{R}^{-1}(\boldsymbol{\gamma})(\mathbf{y}-\mathbf{1}_N\,\mu) \tag{7}$$

2.1 Penalized Likelihood Approach

The penalized likelihood of the collected data is defined as follows

$$Q(\mu,\boldsymbol{\gamma}) = -\frac{1}{2}\log\left|\mathbf{R}(\boldsymbol{\gamma})\right|^{-1/2} - \frac{1}{2}(\mathbf{y}-\mathbf{1}_N\,\mu)^{\mathrm{T}}\,\mathbf{R}^{-1}(\boldsymbol{\gamma})(\mathbf{y}-\mathbf{1}_N\,\mu) - \sum_{k=1}^{d} p_\lambda(\gamma_k) \tag{8}$$

where $p_\lambda(.)$ is a given nonnegative penalty function with a regularization parameter λ. Maximizing the penalized likelihood yields a penalized likelihood estimate $\hat{\mu}$ and $\hat{\boldsymbol{\gamma}}$ for μ and $\boldsymbol{\gamma}$, respectively. For any $\mathbf{x}$, denote

$$\mathbf{b}(\mathbf{x}) = \left(\hat{r}(\mathbf{x},\mathbf{x}_1), \cdots, \hat{r}(\mathbf{x},\mathbf{x}_N)\right) \tag{9a}$$

where

$$\hat{r}(\mathbf{x},\mathbf{x}_1) = \hat{\sigma}^2 \exp\left\{-\sum_{k=1}^{d}\hat{\theta}_k \left|x_k - x_{ik}\right|^q\right\} \tag{9b}$$

The response prediction can be calculated by the best linear unbiased predictor (BLUP)

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{b}(\mathbf{x})\mathbf{R}^{-1}(\hat{\boldsymbol{\gamma}})(\mathbf{y}-\mathbf{1}_N\,\hat{\mu}) \tag{10}$$

with estimated variance

$$\widehat{\mathrm{Var}}\left\{\hat{y}(\mathbf{x})\right\} = \hat{\sigma}^2 - \mathbf{b}(\mathbf{x})\mathbf{R}^{-1}(\hat{\boldsymbol{\gamma}})\mathbf{b}^{\mathrm{T}}(\mathbf{x}) \tag{11}$$

2.2 Relationship of Kriging Model to Other Models

The best linear unbiased predictor (BLUP) in (10) may be viewed as gaussian basis function expansions as follows. Denote,

$$\hat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\left(\hat{\boldsymbol{\gamma}}\right)\left(\mathbf{y} - \mathbf{1}_N\,\hat{\mu}\right) \tag{12}$$

The best linear unbiased predictor for the response variable at input variable $\mathbf{x}$ is

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{b}(\mathbf{x})\hat{\boldsymbol{\beta}} \tag{13}$$

where $\mathbf{b}(\mathbf{x})$, defined by (9), consists of Gaussian basis functions at the sample data points. The model expression in (13) provides a linkage to other modeling techniques such as Kernel Regression and Radial Basis Function (RBF).

RBF uses kernel functions as basis functions and leads to similar model in (13) where the kernel, $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_M(\mathbf{x}))$ and $M \le N$, is in the following *radial* form (i.e., $\theta_1 = \theta_2 = \ldots = \theta$).

$$b_j(\mathbf{x}) = K\left(\theta\left\|\mathbf{x} - \boldsymbol{\xi}_j\right\|\right) \tag{14}$$

where $\|.\|$ is the Euclidean norm. Since the unit can be different in each coordinate, the predictor is usually standardized into unit standard deviation prior to fitting the model. The popular choice for the kernel $K(.)$ is the standard Gaussian density function. Each basis function in (14) is indexed by a "*prototype*" location parameter ξ_j. A popular adhoc practice is to locate ξ_j using clustering method (Bishop, 1995). When the prototype location parameters are set to be equal to the sample data points $\xi_j = \mathbf{x}_j$, $j = 1, \ldots, N$, the RBF model resembles the kernel regression and the Gaussian kriging model. Thus, penalized likelihood in (8) can be used to estimate the model parameters.

3. PENALTY FUNCTION AND ALGORITHM FOR PARAMETER ESTIMATION

In this section, we propose the *Smoothly Clipped Absolute Deviation* (SCAD) penalty as the appropriate choice penalty function $p_\lambda(.)$ and the choice of regularization parameter λ needed for the penalized likelihood in (8). A practical algorithm using the Fisher Scoring approach is employed to estimate the model parameters (9)-(11). The relationship of the SCAD penalty function to L_q penalty functions is briefly discussed. The performance comparison of these penalty functions is discussed in Section 4 where an engineering example is used to illustrate the advantage of the proposed method.

3.1. Selection of a Penalty Function

To facilitate the need of different purposes of model selection, many authors have considered the issue of selection of penalty functions. In the context of linear regression, penalized least squares with L_2 penalty, $p_\lambda(|\beta|) = ½\ \lambda|\beta|^2$, leads to a ridge regression. While the penalized least squares with L_1 penalty, defined by $p_\lambda(|\beta|) = \lambda|\beta|$, corresponds to LASSO (Tibshirani, 1996). Fan and Li (2001) proposed a new penalty function, called *Smoothly Clipped Absolute Deviation* (SCAD) penalty. The first derivative of SCAD is defined by

$$p'_\lambda(\theta) = \lambda\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)} I(\theta > \lambda)\} \tag{15}$$

for some $a > 2$, $\theta > 0$, and $p_\lambda(0) = 0$. This penalty function involves two unknown parameters λ and a. Justifying from a Bayesian statistical point of view, Fan and Li (2001) suggested using $a = 3.7$. The Bayes risk cannot be reduced much with other choices of a and simultaneous selection of a and λ does not have any significant improvements from our experience. For a comparison of the shape of the three penalty functions aforementioned, we plot the penalty functions of L_1, L_2 and SCAD in Figure 3.

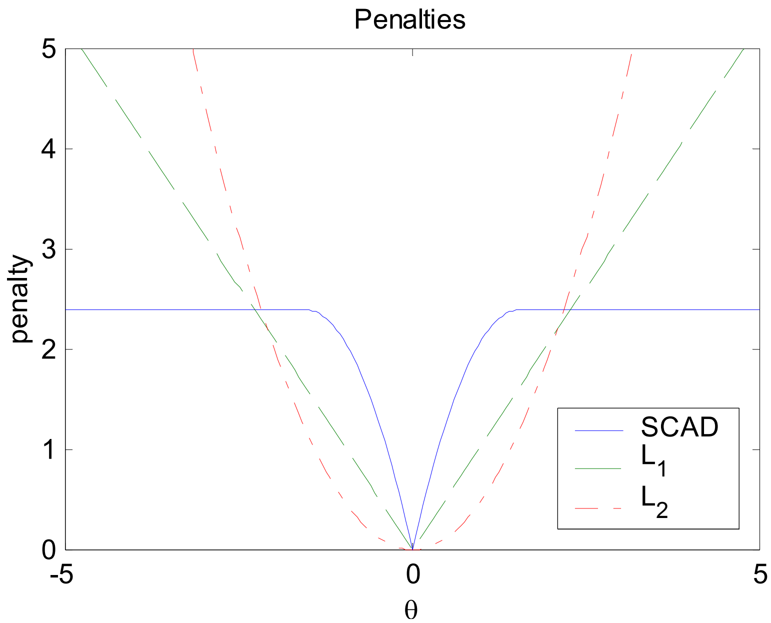

Figure 3. Plot of penalty functions.

3.2. Fisher Scoring Algorithm

Welch et al. (1992) uses stepwise algorithm employing the downhill simplex method to maximize the likelihood function in (7) to sequentially estimate the Gaussian kriging parameters. Here, we use a computationally more efficient gradient-based optimization technique to estimate the parameters. The expression of gradient and Hessian matrix of the penalized likelihood function in (8) is given in the Appendix. Using the first and second order derivative information, one may directly employ Newton-Raphson algorithm to optimize the penalized likelihood. In this paper, Fisher scoring algorithm will be employed to find the solution of the penalized likelihood because of its simplicity and stability. Notice that $E\{\partial^2 l(\mu,\gamma)/\partial\mu\,\partial\gamma\} = 0$ (see

Appendix for the expression of $\partial^2 l(\mu,\gamma)/\partial\mu\,\partial\gamma$). Therefore, the updates of $\hat{\mu}$ and $\hat{\boldsymbol{\gamma}}$ are obtained by solving separate equations. For a given value $\gamma^{(k)}$ of the current step γ at the k-step, the new value $\hat{\mu}$ is obtained by

$$\mu^{(k+1)} = \left(\mathbf{1}_N^{\mathrm{T}}\mathbf{R}^{-1}\left(\boldsymbol{\gamma}^{(k)}\right)\mathbf{1}_N\right)^{-1}\mathbf{1}_N^{\mathrm{T}}\mathbf{R}^{-1}\left(\boldsymbol{\gamma}^{(k)}\right)\mathbf{y} \tag{16a}$$

and

$$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} + \left\{I_{22}\left(\boldsymbol{\gamma}^{(k)}\right) + \Sigma\left(\boldsymbol{\gamma}^{(k)}\right)\right\}^{-1}\partial Q\left(\mu^{(k)},\boldsymbol{\gamma}^{(k)}\right)\Big/\partial\boldsymbol{\gamma} \tag{16b}$$

where $I_{22} = -E\{\partial^2 l(\mu,\gamma)/\partial\gamma\,\partial\gamma\}$, and $\Sigma(\boldsymbol{\gamma}) = \mathrm{diag}\left\{p_\lambda''(\gamma_1),\ldots,p_\lambda''(\gamma_d),0\right\}$, a $(d+1)\times(d+1)$ diagonal matrix.

3.3. Choice of Regularization Parameter

Since Gaussian kriging gives us an exact fit at the sample point $\mathbf{x}$, the residual at each sample point is exactly equal to zero. Therefore, generalized cross-validation (GCV) cannot be used to choose the regularization parameter λ. In this paper, cross-validation (CV) will be used to select the regularization parameter. V-fold CV will be implemented in this paper, and for a given λ, the V-fold CV score is computed in the following way:

1. The data set $\mathcal{D} = \{(\mathbf{x}_i, y_i): i = 1, \ldots, N\}$ are split into V subsets $\mathcal{D}_1, \ldots, \mathcal{D}_V$.

2. For $v = 1, \ldots, V$, let $\mathcal{D}^{(-v)} = \mathcal{D} - \mathcal{D}_v$, and use data $\mathcal{D}^{(-v)}$ to form a predictor $\hat{y}^{(-v)}(\mathbf{x})$.

3. Compute cross-validation score:

$$CV(\lambda) = \sum_v \sum_{(\mathbf{x}_i, y_i)\in\mathcal{D}_v}\left\{y_i - \hat{y}^{(-v)}(\mathbf{x}_i)\right\}^2 \tag{17}$$

For a given set of λ values $S = \{\lambda_1, \ldots, \lambda_K\}$, we choose

$$\hat{\lambda} = \underset{\lambda_k \in S}{\operatorname{argmin}} \; CV\left(\lambda_k\right) \tag{18}$$

In the literature, leave-one-out cross-validation corresponds to *N*-fold cross-validation.

3.4. Computing Algorithm for the Proposed Procedure

We summarize the above procedures in the following algorithm.

1. Choose a grid point set S for λ, say, $\{\lambda_1, \ldots, \lambda_K\}$ and let $i = 1$
2. With λ_i, use the Fisher score algorithm to compute μ and γ
3. Compute the cross-validation score $CV(\lambda_i)$. Let $i = i + 1$
4. Repeat steps 2-3 until all K grid points are exhausted.
5. The final estimator for μ and γ is the one that has the lowest CV score.

4. APPLICATION: PISTON SLAP NOISE

Total vehicle customer satisfaction highly depends on the level of satisfaction a customer has with the vehicle's engine. The Noise, Vibration and Harshness (NVH) characteristics of the vehicle and the engine is one of the critical elements of customer dissatisfaction. Piston slap is an unwanted engine noise that is the result of piston secondary motion. De Luca and Gerges (1996) gave a comprehensive review of the piston slap mechanism and experimental piston slap analysis including noise source analysis and parameters influencing piston slap. Since then, with the advent of faster, more powerful computers, much of the piston slap study has shifted from experimental analysis to analytical analysis for both the power cylinder design phase and for piston noise troubleshooting. Thus, it is desirable to have an analytical model to describe the relationship between the piston slap noise and its covariates, such as, piston skirt length, profile and ovality.

We first give a brief description of this study. A detailed and thorough description of this study can be found in Hoffman et al. (2002). Piston slap as an unwanted engine noise is a result of piston secondary motion, that is, the departure of the piston from the nominal motion prescribed by the slider crank mechanism. The secondary motion is caused by a combination of transient forces and moments acting on the piston during engine operation and the presence of clearances between the piston and the cylinder liner. This combination results in both a lateral movement of the piston within the cylinder and a rotation of the piston about the piston pin, and it causes the piston to impact the cylinder wall at regular intervals. These impacts may result in the objectionable engine noise known as piston slap.

For this study, the power cylinder system was modeled using the multi-body dynamics code ADAMS/Flex that also includes a finite element model. The piston, wrist pin and connecting rod were modeled as flexible bodies, where flexibility is introduced via a model superposition. Boundary conditions for the flexible bodies are included via a Craig-Bampton component mode synthesis. The crankshaft is modeled as a rigid body rotating with a constant angular velocity. In addition, variation in clearance due to cylinder bore distortion and piston skirt profile and ovality will be included in the analysis.

We take the piston slap noise as the output variable, and set clearance between the piston and the cylinder liner (x_1), Location of Peak Pressure (x_2), skirt length (x_3), skirt profile (x_4), skirt ovality (x_5) and pin offset (x_6) as the input variables. Since each computer experiment requires intensive computational resources, uniform design (Fang, 1980) was employed to construct a design of computer experiment with 12 runs. The *Centered*-L_2 discrepancy uniformity criterion

(Fang et al., 2000) was used and optimized using Stochastic Evolutionary algorithm (Jin et al., 2003). A review on uniform design and its applications can be found in (Fang et al., 2000). The collected data are displayed in Table 1. The ultimate goal of the study is to perform robust design optimization (Hoffman et al., 2003) to desensitize the piston slap noise from the source of variability (e.g., clearance variation). To accomplish this goal, the availability of a good metamodel is a necessity. A Gaussian kriging model is employed to construct a metamodel as an approximation to the computationally intensive analytical model. In this discussion, we only focus on the development of the metamodel. Interested readers should consult Hoffman et al. (2003) and Du et al. (2003) for the probabilistic design optimization study.

Table 1. Piston slap noise data.

RUN#	X_1	X_2	X_3	X_4	X_5	X_6	**NOISE (dB)**
1	71	16.8	21	2	1	0.98	56.75
2	15	15.6	21.8	1	2	1.3	57.65
3	29	14.4	25	2	1	1.14	53.97
4	85	14.4	21.8	2	3	0.66	58.77
5	29	12	21	3	2	0.82	56.34
6	57	12	23.4	1	3	0.98	56.85
7	85	13.2	24.2	3	2	1.3	56.68
8	71	18	25	1	2	0.82	58.45
9	43	18	22.6	3	3	1.14	55.5
10	15	16.8	24.2	2	3	0.5	52.77
11	43	13.2	22.6	1	1	0.5	57.36
12	57	15.6	23.4	3	1	0.66	59.64

4.1. Preliminary Analysis: Radial Basis Function Approach

To quickly gain a rough picture on the logarithm of profile likelihood function (log-likelihood, for short), we use a radial basis function approach in (14) by intuitively set $\theta_j = \theta/\sigma_j$, where σ_j stands for the standard deviation of the j-th component of $\mathbf{x}_i$, i = 1, …, N. Such choice of θ_j allows us to plot the log-likelihood function $l(\theta)$ against θ. Plots of the log-likelihood function, and the penalized log-likelihood functions with the SCAD, the L_1 and the L_2 penalties, where $\lambda = 0.2275$ (=$1/2\sqrt{\log(N)/N}$) and N = 12, are depicted in Figure 4, from which we can see that the log-likelihood function is flat when the log-likelihood function near its optimum is flat.

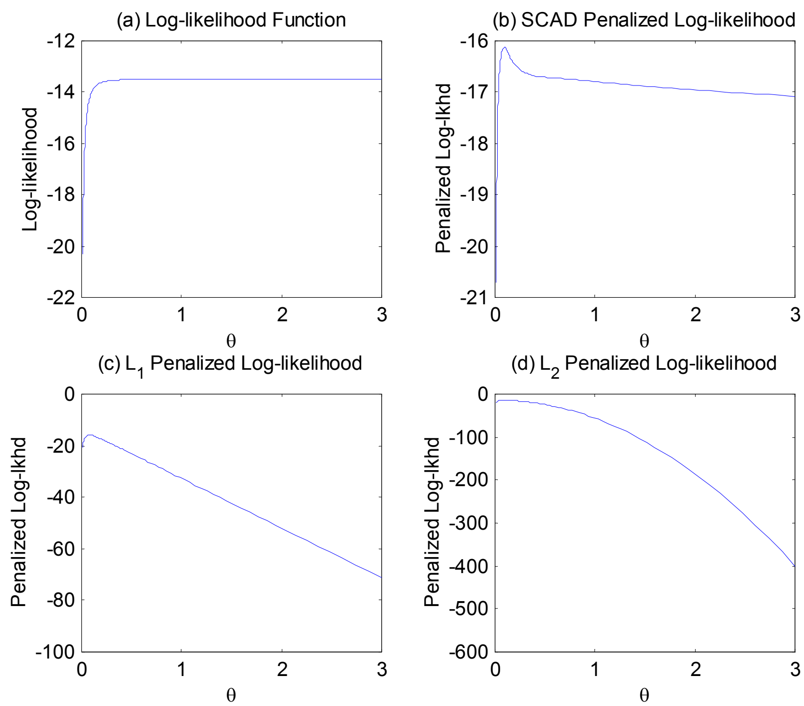

Figure 4. Log-likelihood and penalized log-likelihood when N = 12. (a) is log-likelihood function. (b), (c) and (d) are penalized log-likelihood functions with the SCAD, L_1 and L_2 penalty, respectively.

This flat likelihood function creates the same problem as that in the simple sinusoidal function example (1) discussed in Section 1 when the sample size equals 6. On the contrary, all of the three penalized log-likelihood functions near its optimum are not flat. Although the shape of the corresponding penalized likelihood functions looks very different, their resulting penalized maximum likelihood estimate for θ under the constraint $\theta_j = \theta/\sigma_j$ and $\theta_j \geq 0$ are very close. From the shape of penalized log-likelihood functions, the resulting estimate of penalized likelihood with the SCAD penalty may be more efficient than the other two. This preliminary analysis not only gives us a rough picture of log-likelihood function and the penalized likelihood function, but also provides us with a good initial value in the implementation of Fisher scoring algorithm. We will further demonstrate in next section that the resulting penalized likelihood estimate with the form $\theta_j = \theta/\sigma_j$ also results in a good prediction rule for the output variable.

4.2. Data Analysis via Penalized Gaussian Kriging

The Fisher score algorithm with the initial value obtained in the last section was applied for the data. The leave-one-out cross-validation procedure was used to estimate the tuning parameter λ. The resulting estimate of λ equals 0.1100, 0.1300 and 0.06 for the SCAD, the L_1 and the L_2 penalties, respectively. The resulting estimate of μ, σ^2, θ_j's is depicted in Table 2. The four estimates for μ are very close. But the four penalized likelihood estimates for σ^2 and θ_j's are quite different.

Table 2. Penalized Maximum Likelihood Estimate

PARAMETER	MLE	SCAD	L_1	L_2
$\hat{\mu}$	56.7275	56.2596	56.5177	56.5321
$\hat{\sigma}^2$	3.4844	4.1170	3.6321	3.4854
$\hat{\theta}_1$	0.1397	8.23E-04	1.67E-03	3.78E-03
$\hat{\theta}_2$	1.6300	1.86E-07	1.42E-04	2.43E-02
$\hat{\theta}_3$	2.4451	4.27E-02	0.5779	0.2909
$\hat{\theta}_4$	4.0914	5.61E-07	2.02E-04	3.26E-02
$\hat{\theta}_5$	4.0914	3.03E-06	0.1501	9.80E-02
$\hat{\theta}_6$	12.2253	4.6269	1.48E-02	0.2590

To assess the performance of the penalized Gaussian kriging approach, we conducted another computer experiment with 100 runs. *Median of Absolute Residuals* (MAR) is defined as

$$\mathrm{MAR} = \mathrm{median}\left\{\left|y(\mathbf{x}_i) - \hat{y}(\mathbf{x}_i)\right| : i = 1, \ldots, 100\right\} \tag{19}$$

Equation (19) is employed to measure how well the prediction performs. The MAR for the ordinary kriging method is 1.3375, and MARs equals to 1.0588, 1.4638 and 1.3114 for the penalized kriging method with the SCAD, the L_1 and the L_2 penalties, respectively. The box-plots for the absolute residuals are depicted in Figure 5. The prediction obtained by the penalized kriging with the SCAD penalty outperforms the other three predictions. The performance of the ordinary kriging method is similar to that of penalized kriging method with the L_2 penalty.

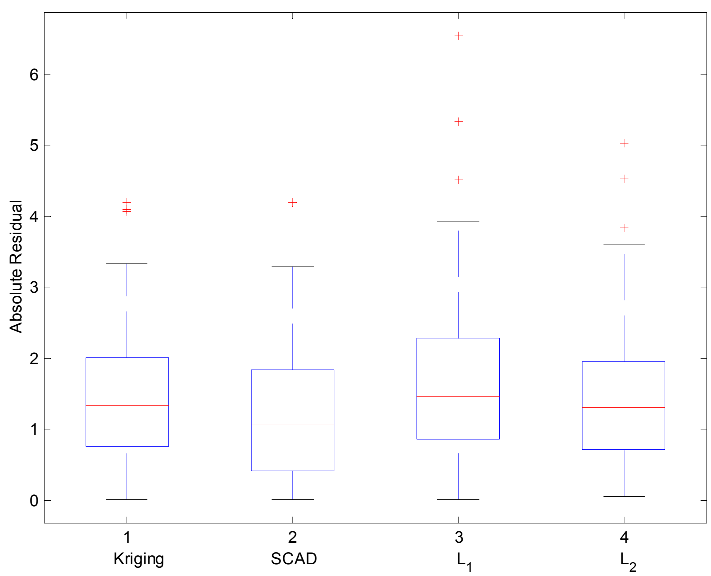

Figure 5. Box-plots of Absolute Residuals.

To understand behavior of the penalized kriging method when the sample size is moderate, we apply the penalized kriging method for the new sample with 100 runs. Again, let $\theta_j = \theta/\sigma_j$, and plot the log-likelihood against θ in Figure 6, from which we can see that the shape of the log-likelihood function is the same as that of the penalized log-likelihood function with the SCAD penalty.

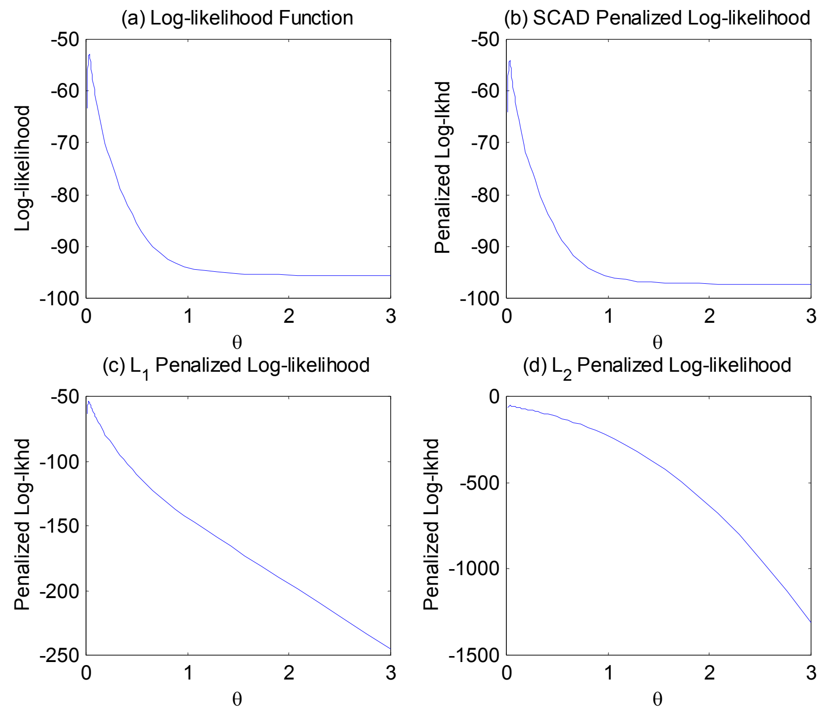

Figure 6. Log-likelihood and penalized log-likelihood when N = 100. (a) is log-likelihood function, (b), (c) and (d) are penalized log-likelihood functions with the SCAD, L_1 and L_2 penalty, respectively.

We further compute the maximum likelihood estimate for all of parameters θ_j. Based on 5-fold cross-validation, the selected λ equals 0.18, 0.105 and 0.18 for the SCAD, L_1 and L_2 penalty. The resulting estimate is listed in Table 3, from which we found that all of these estimate are very close as expected.

4.3. Sensitivity Analysis of Regularization Parameter

Now, we examine how sensitive the penalized likelihood estimate for θ_j to change of the tuning parameter λ and change of the output variables. In this section, we focus on the penalized

Table 3. Penalized Maximum Likelihood Estimate

Parameter	MLE	SCAD	L_1	L_2
$\hat{\theta}_1$	0.4514E-4	0.3971E-4	0.3943E-4	0.5858E-4
$\hat{\theta}_2$	0.5634E-3	0.5192E-3	0.5204E-3	0.6519E-3
$\hat{\theta}_3$	0.3150E-5	0.2602E-5	0.2618E-5	0.4261E-5
$\hat{\theta}_4$	0.2880	0.2752	0.2765	0.3003
$\hat{\theta}_5$	0.2939E-1	0.2593E-1	0.2590E-1	0.3641E-1
$\hat{\theta}_6$	0.1792	0.1515	0.1548	0.2162

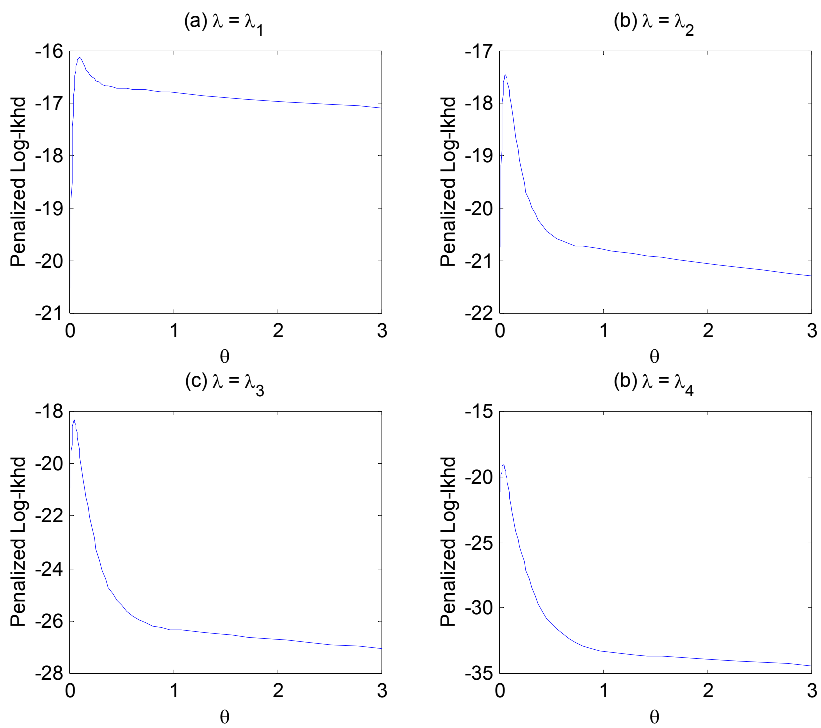

Figure 7. Sensitivity analysis for different choices of $\lambda = \alpha\sqrt{\log(N)/N}$, for $\alpha = 0.5$, 0.75, 1.0, and 1.25, respectively with $N = 12$.

kriging method with the SCAD penalty. We further set $\theta_j = \theta/\sigma_j$. Figure 7 depicts the plots of the log-likelihood function with 4 different choice of $\lambda = 0.2275$, 0.3413, 0.4551 and 0.5688. These values equal $\alpha\sqrt{\log(N)/N}$, for α = 0.5, 0.75, 1.0, and 1.25, respectively with N = 12. From Figure 7, the optimal value for θ is insensitive with respect to the change of λ.

Furthermore, we evaluate the analysis of sensitivity of small change of outputs by adding random perturbation to the output while keeping the input variables the same. Here, we add a small normal noise $N(0, \sigma^2/4)$ to each output, where σ equals the standard deviation of y_i. The log-likelihood functions of two typical data sets are depicted in Figure 8 (a) and (c), from which we can see that the estimate of θ may dramatically change from Figure 8 (a) to (c). The penalized log-likelihood function with the SCAD penalty and λ = 0.2275 for the typical two data sets are displayed in Figure 8 (b) and (d), which shows that the resulting estimate for θ remains the same. This clearly shows that the penalized kriging method with the SCAD penalty is non-sensitive to a small change of output variable.

Now, we further demonstrate the penalized kriging method with the SCAD penalty is also non-sensitive to large change of output variable. To this end, we add a large normal error $N(0, b\sigma^2)$ where b = 0.25, 1, 2, and 3, to each output. The penalized log-likelihood function with the SCAD penalty and λ = 0.2275 for a typical data set is depicted in Figure 9, which clearly shows that the penalized SCAD likelihood estimate for θ is robust to the added noise.

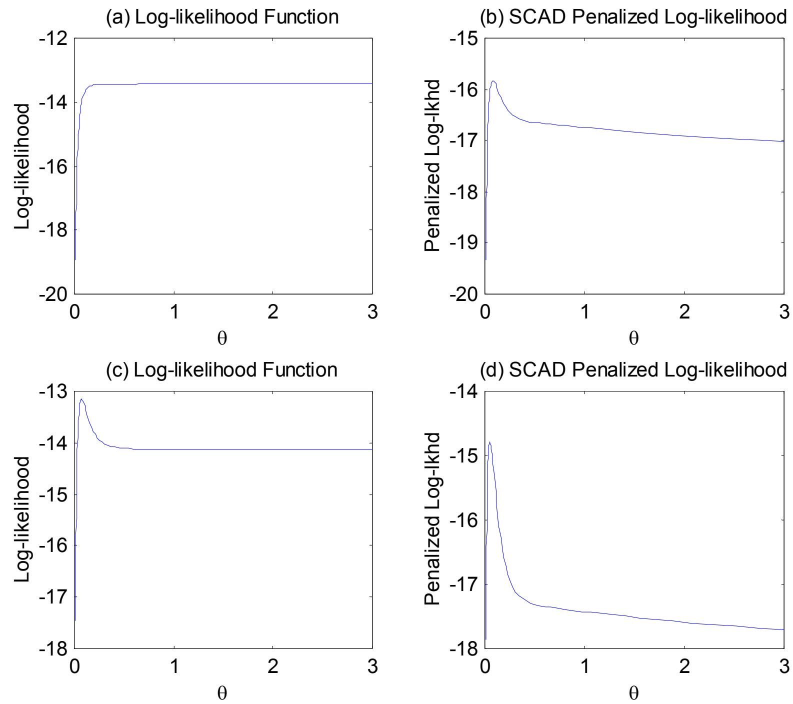

Figure 8. Sensitivity analysis for Gaussian kriging and SCAD penalized kriging method when the outputs were disturbed by adding small normal noise $N(0,\ 0.25\sigma^2)$, where σ equals the standard deviation of the outputs. (a) and (c) are plots of two typical log-likelihood functions of the Guassian kriging method. (b) and (d) are plots of two typical SCAD penalized kriging method.

4.4. Model Interpretation: Main Effects and Interaction Effects

We represent $y(\mathbf{x})$ to be the functional ANOVA form:

$$y(\mathbf{x}) = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots + f_{12\cdots d}(x_1, \cdots, x_d) \tag{20}$$

where

$$\int y(\mathbf{x}) = f_0 \tag{21a}$$

$$\int y(\mathbf{x}) \prod_{k \neq i} dx_k = f_0 + \sum_i f_i(x_i) \tag{21b}$$

$$\int y(\mathbf{x}) \prod_{k \neq i,j} dx_k = f_0 + \sum_i f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) \tag{21c}$$

and so on. Thus, $f_i(x_i)$ can be viewed as the main effects, while $f_{ij}(x_i, x_j)$ may be regarded as interaction effects.

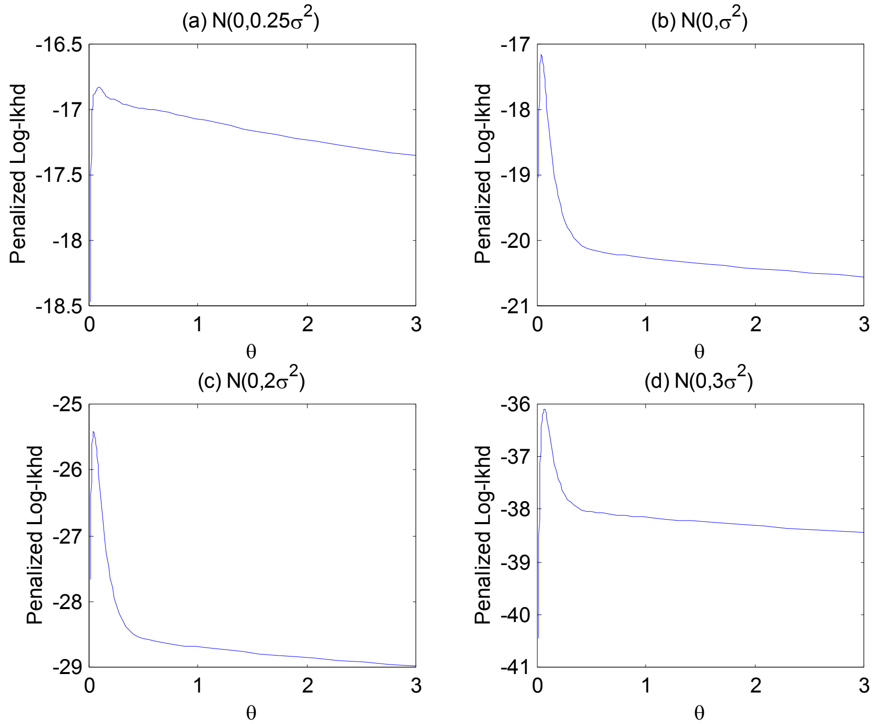

Figure 9. Sensitivity analysis for SCAD penalized kriging method when the outputs were disturbed by adding large normal noise. (a)-(d) are plots of a typical SCAD penalized log-likelihood function, in which σ equals the standard deviation of the outputs.

Extending the definition of indices for ranking importance of input variables for unit cube (Sobol, 2001) to general experimental domain, we define the following indices for ranking importance of input variables:

$$D_0 = \{Vol(\Delta)\}^{-1} \int_{\Delta} \{y(\mathbf{x}) - f_0\}^2 \, d\mathbf{x} \tag{22}$$

and

$$D_{i_1 \dots i_s} = \{Vol(\Delta_{i_1 \dots i_s})\}^{-1} \int_{\Delta_{i_1 \dots i_s}} f^2_{i_1 \dots i_s}(x_{i_1}, \dots, x_{i_s}) dx_{i_1} \dots dx_{i_s} \tag{23}$$

where *Vol(A)* stands for volume of integration domain *A*, Δ is the experimental domain, and $\Delta_{i_1 \ldots i_s}$ is the corresponding experimental domain for $x_{i_1},\ldots,x_{i_s}$. The indices in (22) and (23) can be used to define global sensitivity indices (Sobol, 2001). We computed D_0 = 5.2794E-04, D_i and D_{ij}, tabulated in Table 5, in which the diagonal elements correspond to D_i and off-diagonal element to D_{ij}. Table 5 tells us there are strong interaction effects among clearance, skirt length and pin offset. The main effects, $f_i(x_i)$, is depicted in Figure 10, from which, we can see that the clearance has the strongest main effect, and pin offset also has strong main effect. Although Figure 10 shows weak main effects of location of peak pressure, skirt profile and skirt ovality, Table 5 indicates that the interaction effects of these three factors with other ones are stronger than their main effects.

5. CONCLUSION

Smoothly Clipped Absolute Deviation (SCAD) penalized maximum likelihood estimation has been proposed to deal with problematic flat likelihood function in Gaussian kriging model parameter estimation. Practical implementations of the method are provided including Fisher scoring algorithm as well as the choice and sensitivity of regularization parameter. Comparisons to the standard maximum likelihood estimation as well as L_1 and L_2 penalized likelihood are presented using both toy and industrial applications. The method is particularly recommended for constructing a Gaussian kriging metamodel when regular maximum likelihood estimation result is unsatisfactory, a commonly encountered problem when sample size is small due to computationally expensive engineering simulation models.

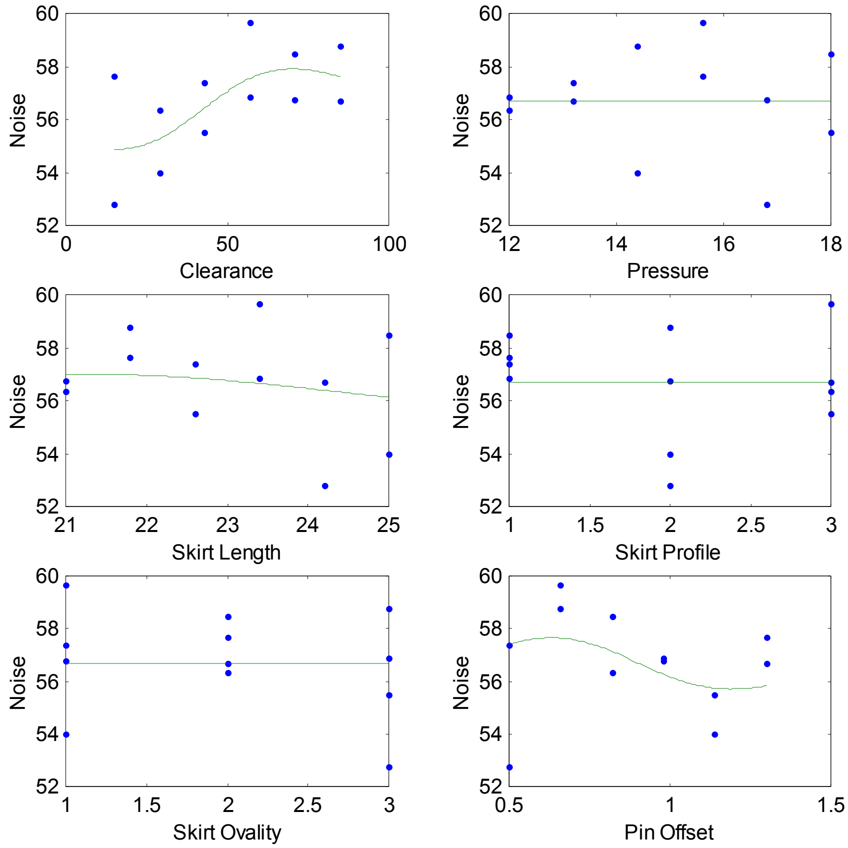

Figure 10. Main effect plots for piston noise. Solid line stands for $f_0 + f_i(x_i)$, where $f_0 = 56.6844$, and dots for the individual observations.

Table 4. Values of D_i and D_{ij}

Factor	Clearance	Press	Skirt Length	Skirt Profile	Skirt Ovality	Pin Offset
Clearance	1.3109					
Press	6.12E-06	2.62E-12				
Skirt Length	0.1109	2.33E-04	7.82E-02			
Skirt Profile	6.12E-06	7.07E-05	2.44E-04	7.05E-12		
Skirt Ovality	6.12E-06	7.07E-05	2.44E-04	7.07E-05	1.38E-11	
Pin Offset	0.7332	3.83E-05	3.31E-03	1.53E-05	6.13E-06	0.5995

APPENDIX

Derivative of $l(\mu, \gamma)$, by some straightforward calculations,

$$\partial l(\mu,\boldsymbol{\gamma})/\partial\mu = -\mathbf{1}_N^{\mathrm{T}}\,\mathbf{R}^{-1}(\boldsymbol{\gamma})\mathbf{e}$$
$$\partial l(\mu,\boldsymbol{\gamma})/\partial\gamma_k = \frac{1}{2}tr\left[\mathbf{R}^{-1}(\boldsymbol{\gamma})\left\{\mathbf{e}\mathbf{e}^{\mathrm{T}} - \mathbf{R}(\boldsymbol{\gamma})\right\}\mathbf{R}^{-1}(\boldsymbol{\gamma})\dot{R}_k(\boldsymbol{\gamma})\right]$$

for $k = 1, \ldots, d+1$, where $\mathbf{e} = \mathbf{y} - \mathbf{1}_N\,\mu,\ \dot{R}_k(\boldsymbol{\gamma}) = \partial\mathbf{R}(\boldsymbol{\gamma})/\partial\gamma_k$

Furthermore, we have

$$\partial^2 l(\mu,\boldsymbol{\gamma})/\partial\mu^2 = \mathbf{1}_N^{\mathrm{T}}\,\mathbf{R}^{-1}(\boldsymbol{\gamma})\mathbf{1}_N$$
$$\partial^2 l(\mu,\boldsymbol{\gamma})/\partial\mu\,\partial\gamma_k = -\mathbf{1}_N^{\mathrm{T}}\,\mathbf{R}^{-1}(\boldsymbol{\gamma})\dot{R}_k(\boldsymbol{\gamma})\mathbf{R}^{-1}(\boldsymbol{\gamma})\mathbf{e}$$
$$\partial^2 l(\mu,\boldsymbol{\gamma})/\partial\gamma_k\,\partial\gamma_s =$$
$$-\frac{1}{2}tr\left[\mathbf{R}^{-1}(\boldsymbol{\gamma})\dot{R}_k(\boldsymbol{\gamma})\mathbf{R}^{-1}(\boldsymbol{\gamma})\left\{2\,\mathbf{e}\mathbf{e}^{\mathrm{T}} - \mathbf{R}(\boldsymbol{\gamma})\right\}\mathbf{R}^{-1}(\boldsymbol{\gamma})\dot{R}_s(\boldsymbol{\gamma})\right]$$
$$+\frac{1}{2}tr\left[\mathbf{R}^{-1}(\boldsymbol{\gamma})\left\{\mathbf{e}\mathbf{e}^{\mathrm{T}} - \mathbf{R}(\boldsymbol{\gamma})\right\}\mathbf{R}^{-1}(\boldsymbol{\gamma})\ddot{R}_{ks}(\boldsymbol{\gamma})\right]$$
$$k, s = 1, \ldots, d+1 \text{ and } \ddot{R}_{ks}(\boldsymbol{\gamma}) = \partial^2\mathbf{R}(\boldsymbol{\gamma})/\partial\gamma_k\,\partial\gamma_s$$

ACKNOWLEDGMENTS

We are thankful to Joseph Stout of Ford Motor Company to allow us to use the data in his project for our case study. Li's research was supported by a NSF grant DMS-0102505.

REFERENCES

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press.

Booker, A.J., Dennis, J.E., Jr., Frank, P.D., Serafini, D.B., Torczon, V., and Trosset, M.W. (1999), "A Rigorous Framework for Optimization of Expensive Function by Surrogates," *Structural Optimization*, **17** (1), 1-13.

De Luca, J. C. and Gerges, S. N. Y., (1996), "Piston Slap Excitation: Literature Review," *SAE Paper* No. 962396.

Du, X, and Chen, W. (2002), “Sequential Optimization and Reliability Assessment method for Efficient Probabilistic Design,” *2002 ASME Design Automation Conference*.

Du, X., Sudjianto, A., and Chen, W. (2003), "An Integrated Framework for Optimization under Uncertainty Using Inverse Reliability Strategy," Submitted to *ASME Journal of Mechanical Design*.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of American Statistical Association*, **96**, 1348-1360.

Fang, K. T. (1980), "The Uniform Design: Application Of Number-Theoretic Methods In Experimental Design," *Acta Math. Appl. Sinica*, **3**, 363-372.

Fang, K.T., D.K.J. Lin, P. Winker and Y. Zhang (2000), "Uniform Design: Theory And Applications," *Technometrics*, **42**, 237-248.

Hoffman, R. M., Sudjianto, A., Du, X. and Stout, J. (2003), "Robust Piston Design And Optimization Using Piston Secondary Motion Analysis," *SAE Paper*, No. 2003-01-0148.

Jin, R., Chen, W., and Sudjianto, A. (2003), "An Efficient Algorithm for Constructing Optimal Design of Computer Experiments," Submitted to *2003 ASME DETC*.

Jin, R., Chen, W. Simpson, T.W. (2000), "Comparative Studies of Metamodeling Techniques Under Multiple Modeling Criteria," *AIAA-2000-4801*.

Kalagnanam, J.R. and Diwekar, U.M. (1997), "An Efficient Sampling Techniques for Off-line Quality Control," *Technometrics*, **39**(3), 308-319.

Kleijnen, J.P.C. (1987), *Statistical Tools for Simulation Practitioners*, Marcel Decker, N.Y.

Kodiyalam, S., Yang, R-J., Gu, L., and Tho, C-H. (2001), "Large-Scale, Multidisplinary Optimization of Vehicle System in a Scalable, High Performance Computing Environment," *DETC2001/DAC-21082*.

Koehler, J. R. and Owen, a. B. (1996), "Computer Experiments," *Handbook of Statistics*, Vol. 13, S. Ghosh and C. R. Rao, eds., 261-308. Elsevier, Science B. V.

Meckesheimer, M., Barton, R.R., Simpson, T.W., and Booker, A. (2002), "Computationally Inexpensive Metamodel Assessment Strategies," *AIAA Journal*, **40**, 2053-2060.

Tibshirani, R. (1996), "Regression Shrinkage And Selection via the LASSO," *Journal of Royal Statistical Society*, B, **58**, 267-288.

Sack, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1990), "Design and Analysis of Computer Experiments (with discussion)," *Statistical Science*, **4**, 409-435.

Simpson, T.W., Booker, A.J., Ghosh, D., Giunta, A.A., Koch, P.N., Yang, R-J. (2002), "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion," 9th *AIAA/ISSMO Symposium on Multi-disciplinary Analysis and Optimization*, Atlanta, GA, Sept. 2-4, 2002.

Sobol, I.M. (2001), "Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates," *Mathematics and Computers in Simulation*, **55**, 271-280.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T., and Morris, M.D. (1992), "Screening, Predicting, and Computer Experiments," *Technometrics*, **34**(1), 15-25.

Wu, Y.-T. and Wang W. (1998), “Efficient Probabilistic Design by Converting Reliability Constraints to Approximately Equivalent Deterministic Constraints,” *Journal of Integrated Design and Process Sciences*, **2**(4), 13-21.

Ye, K. Q., Li, W., Sudjianto, A. (2000), "Algorithmic Construction of Optimal Symmetric Latin Hypercube Designs," *Journal of Statistical Planning and Inference*, **90**, 145-159.