Editor's Note: This article will be presented at the *Technometrics* invited session at the 2005 Joint Statistical Meetings in Minneapolis, Minnesota, August 2005.

Analysis of Computer Experiments Using Penalized Likelihood in Gaussian Kriging Models

Runze LI

Department of Statistics
Pennsylvania State University
University Park, PA 16802
(*rli@stat.psu.edu*)

Agus SUDJIANTO

Risk Management Quality & Productivity
Bank of America
Charlotte, NC 28255
(*agus.sudjianto@bankofamerica.com*)

Kriging is a popular analysis approach for computer experiments for the purpose of creating a cheap-to-compute "meta-model" as a surrogate to a computationally expensive engineering simulation model. The maximum likelihood approach is used to estimate the parameters in the kriging model. However, the likelihood function near the optimum may be flat in some situations, which leads to maximum likelihood estimates for the parameters in the covariance matrix that have very large variance. To overcome this difficulty, a penalized likelihood approach is proposed for the kriging model. Both theoretical analysis and empirical experience using real world data suggest that 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 because collection of a large sample set is prohibitive. The proposed approach is applied to 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.

KEY WORDS: Computer experiment; Fisher scoring algorithm; Kriging; Meta-model; Penalized likelihood; 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. Although sophisticated engineering computer simulations have become ubiquitous tools for investigating complicated physical phenomena, their effectiveness in supporting timely design decisions in quick-pace product development is often hindered due to their excessive requirements for model preparation, computation, and output postprocessing. The computational requirement increases dramatically when the simulation models are used for 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.

The outer loop is the optimization itself, and the inner loops are probability calculations for the design objective and design constraint. The most challenging issue for implementing probabilistic design is associated with the intensive computational demand of this double-loop procedure. To deal with this issue, meta-modeling, the "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, Sudjianto, Du, and Stout 2003; Du, Sudjianto, and Chen 2004). Although probabilistic design optimization is beyond the scope of this article, it provides a strong motivation for the approach presented herein, where the ability to construct an accurate meta-model from a small sample size is crucial. For example, it takes 24 hours for every run of computer experiments for the piston slap noise example in Section 4. In such situations, collecting a large sample may be very difficult, and the newly proposed approaches are recommended.

The accuracy of meta-models in representing the original model is influenced both by the experimental designs used (see, e.g., Ye, Li, and Sudjianto 2000) and by the meta-modeling approach itself (Jin, Chen, and Simpson 2000). The design and analysis of computer experiments for meta-modeling has recently received much interest in both the engineering and statistical communities (Welch et al. 1992; Jin et al. 2000; Simpson et al. 2002). Because 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. Koehler and Owen (1996) and Sacks, Welch, Mitchell, and Wynn (1990) provided a detailed reviews on how to scatter computer design points over the experimental domain effectively and how to analyze the deterministic output. Santner, Williams, and Notz (2003) provided a systematic introduction on space-filling designs for computer experiments and a thorough description of prediction methodology for computer experiments. Sacks et al. (1990) advocated modeling the deterministic output as a realization of a stochastic process, and used Gaussian stochastic kriging methods to predict the deterministic outputs. In implementing Gaussian kriging models, one may introduce some parameters in the covariance matrix and use the maximum likelihood approach to construct estimates for the parameters.

Although the Gaussian kriging method is useful and popular in practice (e.g., Booker et al. 1999; Jin et al. 2000; Kodiyalam, Yang, Gu, and Tho 2001; Meckesheimer, Barton, Simpson, and Booker 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

TECHNOMETRICS, MAY 2005, VOL. 47, NO. 2
DOI 10.1198/004017004000000671

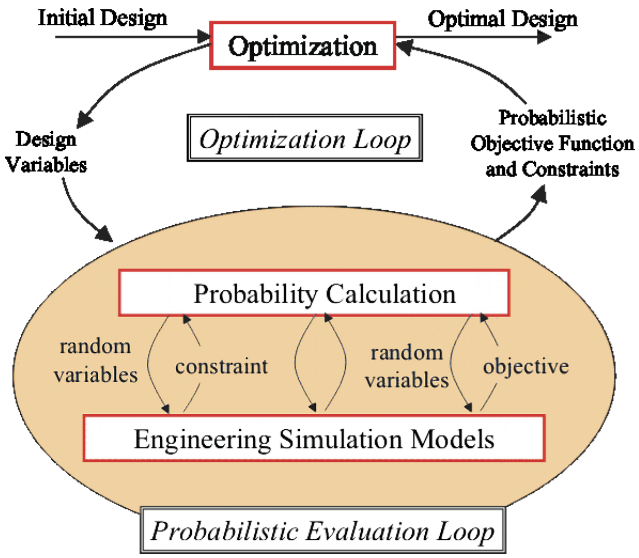

Figure 1. Double-Loop Procedure in Probabilistic Design.

estimates (MLEs) for the parameters in the covariance matrix may have very large variance, because the likelihood function near the optimum is flat. We demonstrate this problem in the following simple example.

Consider the one-dimensional function

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

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

$$y(x) = \mu + z(x),$$

where $z(x)$ is a Gaussian process with mean 0 and covariance

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

For a given θ, the MLE for μ and σ^2 can be easily computed. We can further compute the profile likelihood function $\ell(\theta)$, which equals the maximum of the likelihood function over μ and σ^2 for any given θ. The corresponding logarithm of the profile likelihood (log-likelihood, for short) function $\ell(\theta)$ versus θ is depicted in Figure 2(a), from which we can see that the likelihood function achieves its maximum at $\theta = 3$ and 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.

As a natural alternative approach, one may use the restricted maximum likelihood (REML) (Patterson and Thompson 1971) method to estimate the parameters involving a covariance matrix. (REML is also called "residual" or "modified" maximum likelihood in the literature.) Let $\mathbf{y} = (y_1, \ldots, y_N)$ consist of all responses, and let $\mathbf{C}(\theta)$ be the $N \times N$ correlation matrix whose (i,j)th element is $\exp\{-\theta|x_i - x_j|^2\}$. The logarithm of the restricted likelihood function for this example is

$$\begin{aligned} &\frac{n}{2}\log(2\pi) - \frac{(n-1)}{2}\log\sigma^2 \\ &\quad - \frac{1}{2}\log|\mathbf{C}(\theta)| - \frac{1}{2}\log|\mathbf{1}_N^T\mathbf{C}^{-1}(\theta)\mathbf{1}_N| \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{1}_N\mu)^T\mathbf{C}^{-1}(\theta)(\mathbf{y} - \mathbf{1}_N\mu), \end{aligned}$$

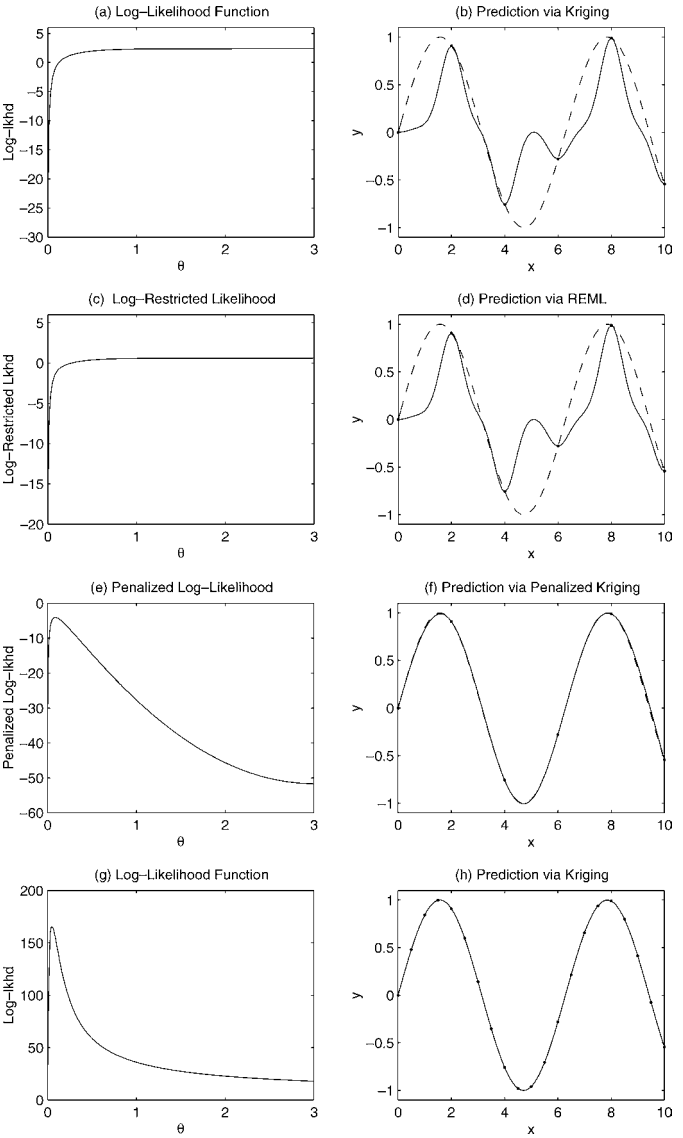

Figure 2. Kriging and Penalized Kriging. (a) and (g) The log-likelihood functions for kriging with sample size N = 6 and 21. (b) and (h) Prediction via kriging with sample size N = 6 and 21. (c) The log-restricted likelihood function for kriging with N = 6. (d) The prediction via kriging with N = 6 using the REML method. (e) The penalized log-likelihood function for kriging with N = 6. (f) The prediction via penalized kriging with N = 6. In (b), (d), (f), and (h), the solid line represents prediction, the dashed line represents the true curve, and the dots represent prediction at the sample datum points.

where $\mathbf{1}_N$ is an N-dimensional vector with all elements equal to 1. The corresponding logarithm of the profile restricted likelihood function versus θ is depicted in Figure 2(c), from which we can see that the shape of the profile restricted likelihood function in this example is the same as that of Figure 2(a). The profile restricted likelihood function achieves its maximum at $\theta = 3$ and becomes almost flat for $\theta \geq 1$. The prediction based on the REML is displayed in Figure 2(d). The prediction is the same as that of Figure 2(b), because $\hat{\theta} = 3$, the same as that obtained by ordinary likelihood approach. As shown in this example, the prediction based on REML also becomes very erratic when x is not equal to the sample data.

To avoid such erratic behavior, we consider a penalized likelihood approach, which we describe in detail in Section 2. A pe-

nalized log-likelihood function with the SCAD penalty (see Sec. 3 for a definition of SCAD) is depicted in Figure 2(e), and its corresponding prediction is displayed in Figure 2(f). Figure 2(e) clearly shows that the penalized likelihood function reaches its maximum at $\theta = .091$ and is not flat around the optimum. The shape of the penalized log-likelihood function implies that the resulting estimate for θ has smaller variance (see Sec. 2.2 for more discussion). Figure 2(f) demonstrates that the predicted and the true curve are almost identical. Although REML may be viewed as a kind of penalized likelihood, the motivation of REML is different from that of our penalized likelihood. For example, the goal of our penalized likelihood is to reduce variance of the resulting estimate of θ at the expense of introducing a small bias (see Sec. 2.2 for theoretic analysis); however, the goal of REML is to produce an unbiased estimate by paying a price of increasing variance of the resulting estimate. This can be easily seen from the REML estimate of variance of random error (σ^2) for the ordinary multiple linear regression model with independent and identically distributed random error.

To demonstrate the effect of sample data on the likelihood function and to properly predict at unsampled points, we consider a slightly larger sample, $x = 0, .5, 1, \ldots, 10$. The corresponding likelihood function and the prediction are shown in Figure 2(g). The likelihood function achieves its maximum at $\theta = .051$. Comparing Figures 2(e) and 2(g), we see that the locations of the MLE and the penalized MLE are very close. Furthermore, Figure 2(h) confirms that the corresponding prediction yielded by the penalized kriging method with $N = 6$ comparable to the prediction obtained by maximum likelihood with $N = 21$.

In this article we propose a new approach via penalized likelihood to fit a Gaussian kriging model to 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 the Fisher scoring algorithm. Furthermore, we propose a method for choosing the regularization parameter involved in the penalty function. We summarize the proposed approach as an easy-to-follow algorithm. We demonstrate the benefit of our proposed method 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 Δ, and that $y_i = y(\mathbf{x}_i)$ are sampled from the model

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

where $z(\mathbf{x}_i)$ is a Gaussian process with mean 0 and covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$,

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

where $\theta_k \ge 0$. Although one may estimate q in practice, we take $q = 2$ throughout the article, because the computer model is known to be smooth. Let $\boldsymbol{\gamma} = (\theta_1, \ldots, \theta_d, \sigma^2)^T$ and define $\mathbf{R}(\boldsymbol{\gamma})$ to be the $N \times N$ matrix with the (i, j)th element $r(\mathbf{x}_i, \mathbf{x}_j)$. Thus the density of $\mathbf{y} = (y_1, \ldots, y_N)^T$ is

$$f(\mathbf{y}) = (2\pi)^{-N/2} |\mathbf{R}(\boldsymbol{\gamma})|^{-1/2} \times \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{1}_N \mu)^T \mathbf{R}^{-1}(\boldsymbol{\gamma}) (\mathbf{y} - \mathbf{1}_N \mu) \right\}, \quad (3)$$

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

$$\ell(\mu, \boldsymbol{\gamma}) = -\frac{1}{2} \log |\mathbf{R}(\boldsymbol{\gamma})| - \frac{1}{2} (\mathbf{y} - \mathbf{1}_N \mu)^T \mathbf{R}^{-1}(\boldsymbol{\gamma}) (\mathbf{y} - \mathbf{1}_N \mu). \quad (4)$$

2.1 Penalized Likelihood and Kriging

The penalized likelihood of the collected data is defined as

$$Q(\mu, \boldsymbol{\gamma}) = -\frac{1}{2} \log |\mathbf{R}(\boldsymbol{\gamma})| - \frac{1}{2} (\mathbf{y} - \mathbf{1}_N \mu)^T \mathbf{R}^{-1}(\boldsymbol{\gamma}) (\mathbf{y} - \mathbf{1}_N \mu) - N \sum_{k=1}^{d} p_\lambda(\gamma_k), \quad (5)$$

where $p_\lambda(\cdot)$ is a given nonnegative penalty function with a regularization parameter λ. Maximizing the penalized likelihood yields penalized likelihood estimates $\hat{\mu}$ and $\hat{\boldsymbol{\gamma}}$ for μ and $\boldsymbol{\gamma}$. The penalized likelihood may be written in an equivalent form of constrained likelihood. For instance, if we take the penalty function to be the L_1 penalty, namely $p_\lambda(\boldsymbol{\gamma}) = \lambda \boldsymbol{\gamma}$, then maximizing the penalized likelihood (5) is equivalent to maximizing the likelihood function $\ell(\mu, \boldsymbol{\gamma})$ subject to a constraint $\sum_{j=1}^{d} \gamma_j \le s$, where s is a regularization parameter playing the same role as that of λ. (See Tibshirani 1996 for detailed discussion on constrained least squares with the L_1 penalty.) Penalized likelihood is closely related to variable selection criteria, such as the Akaike information criterion (Akaike 1974), the Bayes information criterion (Schwarz 1978), and more recent work (Fan and Li 2001). Furthermore, many smoothing methods, including smoothing splines (Wahba 1990) and penalized splines (Ruppert 2002), can be derived from a penalized likelihood. The penalized likelihood also admits Bayesian interpretations where the penalty term corresponds to a prior on $\boldsymbol{\gamma}$. Bayesian interpolation has a long history. Kimdeldorf and Wahba (1970) and Wahba (1978) established a connection between Bayesian interpolation and smoothing splines. It is well known that smoothing splines are the solution of penalized least squares with a quadratic penalty or penalized likelihood with a quadratic penalty for the regression coefficients when random error is normally distributed. Our penalized likelihood approach is differs from smoothing splines in that we penalize the parameters involved in the correlation matrix rather than the regression coefficients for the mean function. Bayesian interpolation was introduced to model computer experiments by Currin, Mitchell, Morris, and Ylvisaker (1991) who they provided a Bayesian formulation for Gaussian kriging models. For any $\mathbf{x}$, denote

$$\mathbf{b}(\mathbf{x}) = \big(\hat{r}(\mathbf{x}, \mathbf{x}_1), \ldots, \hat{r}(\mathbf{x}, \mathbf{x}_N)\big), \quad (6)$$

where

$$\hat{r}(\mathbf{x}, \mathbf{x}_i) = \hat{\sigma}^2 \exp\left\{ -\sum_{k=1}^{d} \hat{\theta}_k |x_k - x_{ik}|^q \right\}. \tag{7}$$

The predicted response can be calculated by the best linear unbiased predictor,

$$\hat{y}(\mathbf{x}) = \hat{\mu} + \mathbf{b}(\mathbf{x})\mathbf{R}^{-1}(\hat{\boldsymbol{\gamma}})(\mathbf{y} - \mathbf{1}_N\hat{\mu}), \tag{8}$$

with estimated variance

$$\widehat{\text{var}}\{\hat{y}(\mathbf{x})\} = \hat{\sigma}^2 - \mathbf{b}(\mathbf{x})\mathbf{R}^{-1}(\hat{\boldsymbol{\gamma}})\mathbf{b}^T(\mathbf{x}). \tag{9}$$

2.2 Theoretical Aspects

We solve the following equations to find the solution of penalized likelihood:

$$\frac{\partial Q(\mu, \boldsymbol{\theta}, \sigma^2)}{\partial \mu} = 0, \tag{10}$$

$$\frac{\partial Q(\mu, \boldsymbol{\theta}, \sigma^2)}{\partial \sigma^2} = 0, \tag{11}$$

and

$$\frac{\partial Q(\mu, \boldsymbol{\theta}, \sigma^2)}{\partial \theta_j} = 0, \quad \text{for } j = 1, \ldots, d. \tag{12}$$

From (10) and (11), we have

$$\hat{\mu} = \mathbf{1}_N^T\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{y}/\mathbf{1}_N^T\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{1}_N \tag{13}$$

and

$$\hat{\sigma}^2 = \frac{1}{N}(\mathbf{y} - \mathbf{1}_N\hat{\mu})^T\mathbf{C}^{-1}(\boldsymbol{\theta})(\mathbf{y} - \mathbf{1}_N\hat{\mu}), \tag{14}$$

where $\mathbf{C}(\boldsymbol{\theta}) = \sigma^{-2}\mathbf{R}(\boldsymbol{\gamma})$. Thus, for given initial values for μ and σ^2, we can use the Newton–Raphson or Fisher scoring algorithm to solve (12) for $\boldsymbol{\theta}$ (see Sec. 3.2 for implementation of the Fisher scoring algorithm). Furthermore, we iteratively update the values of μ and σ^2 using (13) and (14), and solve (12) for $\boldsymbol{\theta}$ until the algorithm converges. Thus, when we solve (12) for $\boldsymbol{\theta}$, the values of μ and σ^2 are fixed.

Because our focus in this article is on estimation of θ_j, $j = 1, \ldots, d$, for simplicity of presentation, we fix μ and σ^2, and regard $\ell(\mu, \theta_1, \ldots, \theta_d, \sigma^2)$ as a function $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$ and denote it by $\ell(\boldsymbol{\theta})$, suppressing μ and σ^2. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ denote the MLE of $\boldsymbol{\theta}$, that is,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}).$$

The consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ has been established by Mardia and Marshall (1984); Sweeting (1980) provided more general settings. In this section we assume that the regularity conditions presented by Sweeting (1980) and Mardia and Marshall (1984) are valid.

We first present some geometric interpretations to explain conditions where the ordinary likelihood method does not work well and the penalized likelihood method is needed. Under some regularity conditions, it follows that

$$\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_0) + \boldsymbol{\ell}'(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\boldsymbol{\ell}''(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o_P(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2),$$

where $\boldsymbol{\ell}'(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\boldsymbol{\ell}''(\boldsymbol{\theta}) = \partial^2\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^T$. When $\boldsymbol{\theta}$ is one-dimensional, if $\ell(\boldsymbol{\theta})$ becomes flat for $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_0$ as shown in Figure 2(a), then $\boldsymbol{\ell}'(\boldsymbol{\theta}) \approx \mathbf{0}$ and $\boldsymbol{\ell}''(\boldsymbol{\theta}_0) \approx \mathbf{0}$. This indicates that the variance of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is very large as it approximately equals the inverse of $-E\{\boldsymbol{\ell}''(\boldsymbol{\theta}_0)\}$ (see more discussion later). For multidimensional $\boldsymbol{\theta}$, when $\boldsymbol{\ell}''(\boldsymbol{\theta}_0)$ is nearly singular, the variance of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ becomes very large. In such situations, the penalized likelihood estimator may perform better than the MLE. Like $\ell(\boldsymbol{\theta})$, let $Q(\boldsymbol{\theta})$ denote the penalized likelihood function with fixed μ and σ. Under certain regularity conditions,

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}_0) + \{\boldsymbol{\ell}'(\boldsymbol{\theta}_0) - NP_1(\boldsymbol{\theta}_0)\}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\{\boldsymbol{\ell}''(\boldsymbol{\theta}_0) - N\mathbf{P}_2(\boldsymbol{\theta}_0)\}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + o_P(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2),$$

where $\mathbf{P}_1(\boldsymbol{\theta}) = (p'_\lambda(\theta_{10}), \ldots, p'_\lambda(\theta_{d0}))^T$ and $\mathbf{P}_2(\boldsymbol{\theta}) = \text{diag}\{p''_\lambda(\theta_{10}), \ldots, p''_\lambda(\theta_{d0})\}$. Note that $\mathbf{P}_2(\boldsymbol{\theta})$ is a diagonal matrix and plays the same role as that of the ridge matrix in ridge regression, which is used to deal with the problem of collinearity. Thus when $\boldsymbol{\ell}''(\boldsymbol{\theta}_0)$ is near singular and the likelihood function becomes flat around $\boldsymbol{\theta}_0$, the penalized likelihood function produces a more stable solution for $\boldsymbol{\theta}$. (Here "stable" means that the resulting estimate has small variance.)

We next study the asymptotic properties of the penalized likelihood estimate $\hat{\boldsymbol{\theta}}$. Because the asymptotic normality of $\hat{\boldsymbol{\theta}}$ requires the same regularity conditions as those for $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, we start with the asymptotic properties of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. As was shown by Mardia and Marshall (1984), under certain regularity conditions, we have

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \text{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)),$$

where "$\xrightarrow{\mathcal{D}}$" represents convergence in distribution and $\mathbf{I}(\boldsymbol{\theta}_0) = -N^{-1}E\{\boldsymbol{\ell}''(\boldsymbol{\theta}_0)\}$ is the Fisher information matrix. Thus when $\mathbf{I}(\boldsymbol{\theta}_0)$ is near singular, $\hat{\boldsymbol{\theta}}$ has a large covariance matrix. This is the case illustrated in Figure 2(a). Hence the behavior of the resulting prediction becomes erratic. Under certain regularity conditions, we can show, using techniques related to those of Fan and Li (2001), that

$$\sqrt{N}\{\mathbf{I}(\boldsymbol{\theta}_0) + \mathbf{P}_2(\boldsymbol{\theta}_0)\}\big[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 + \{\mathbf{I}(\boldsymbol{\theta}_0) + \mathbf{P}_2(\boldsymbol{\theta}_0)\}^{-1}\mathbf{P}_1(\boldsymbol{\theta}_0)\big] \xrightarrow{\mathcal{D}} \text{N}(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)).$$

Thus the asymptotic bias of $\hat{\boldsymbol{\theta}}$ is $\mathbf{P}_1(\boldsymbol{\theta}_0)$, and the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is

$$\frac{1}{N}\{\mathbf{I}(\boldsymbol{\theta}_0) + \mathbf{P}_2(\boldsymbol{\theta}_0)\}^{-1}\mathbf{I}(\boldsymbol{\theta}_0)\{\mathbf{I}(\boldsymbol{\theta}_0) + \mathbf{P}_2(\boldsymbol{\theta}_0)\}^{-1}.$$

The penalized likelihood approach can significantly reduce the variance of $\hat{\boldsymbol{\theta}}$ when $\mathbf{I}(\boldsymbol{\theta}_0)$ is near singular—but at the expense of introducing bias $\mathbf{P}_1(\boldsymbol{\theta}_0)$. Root-$n$ consistency of $\hat{\boldsymbol{\theta}}$ requires that $\max_j |p'_\lambda(\theta_{j0})| = O(N^{-1/2})$. Furthermore, if $\max_j |p'_\lambda(\theta_{j0})| = o(N^{-1/2})$ and $\max_j |p''_\lambda(\theta_{j0})| = o(N^{-1/2})$, then

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \text{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)),$$

which is the same as that of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. This is a desirable property, because it is well known that the MLE is the most efficient estimate. The result also implies that the penalized likelihood estimate may perform better than the MLE only when the sample

size is small. This result has an important practical implications, because computer experiments can be very time-consuming. For example, it takes 24 hours for every run of computer experiments for piston slap noise example analyzed in Section 4. In such situations, collecting a large sample may be very difficult, and penalized likelihood approaches are recommended.

3. PENALTY FUNCTION AND ALGORITHM FOR PARAMETER ESTIMATION

In this section we propose the smoothly clipped absolute deviation (SCAD) penalty as the appropriate choice of penalty function $p_\lambda(\cdot)$ and the choice of regularization parameter λ needed for the penalized likelihood in (5). A practical algorithm using the Fisher scoring approach is used to estimate the model parameters μ, σ, and $\boldsymbol{\theta}$. The performance comparison of these penalty functions is discussed in Section 4, in which an engineering example is used to illustrate the advantage of the proposed method.

3.1 Selection of a Penalty Function

Because model selection is used for various purposes, many authors have considered the issue of selecting penalty functions. In the context of linear regression, penalized least squares with L_2 penalty, $p_\lambda(|\theta|) = .5\lambda|\theta|^2$, leads to a ridge regression, whereas the penalized least squares with L_1 penalty, defined by $p_\lambda(|\theta|) = \lambda|\theta|$, corresponds to LASSO (Tibshirani 1996). Fan and Li (2001) proposed a new penalty function, the SCAD penalty. The first derivative of SCAD is defined by

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

for some $a > 2$, $\theta > 0$, with $p_\lambda(0) = 0$. This penalty function involves two unknown parameters λ and a. As suggested by Fan and Li (2001), we set $a = 3.7$ throughout the article. As demonstrated by Fan and Li (2001), the performance cannot be significantly improved with a selected by data-driven methods, such as cross-validation. Furthermore, the data-driven method can be very computationally extensive, because one needs to search for an optimal pair (λ, a) over a two-dimensional grid of points. The shapes of the three penalty functions (L_1, L_2, and SCAD) are shown in Figure 3.

As discussed in Section 2.2, if $\max_j |p'_\lambda(\theta_{j0})| = o(N^{-1/2})$ and $\max_j |p''_\lambda(\theta_{j0})| = o(N^{-1/2})$, then

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathrm{N}(\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)), \qquad (16)$$

which is the same as that of $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$. For the L_2 penalty, $p'_\lambda(\boldsymbol{\theta}) = \lambda\boldsymbol{\theta}$ and $p''_\lambda(\boldsymbol{\theta}) = \lambda$. Thus, when $\lambda = o(N^{-1/2})$ for the L_2 penalty, then, under certain regularity conditions, (16) holds. For the L_1 penalty, $p'_\lambda(\boldsymbol{\theta}) = \lambda$ and $p''_\lambda(\boldsymbol{\theta}) = 0$. Hence (16) requires that $\lambda = o(N^{-1/2})$. As to the SCAD penalty, when $\lambda \to 0$, $\max_j |p'_\lambda(\theta_{j0})| \to 0$ and $\max_j |p''_\lambda(\theta_{j0})| \to 0$ for any given $\theta_{j0} > 0$. This implies that if $\lambda = o(1)$ for the SCAD penalty, then (16) holds.

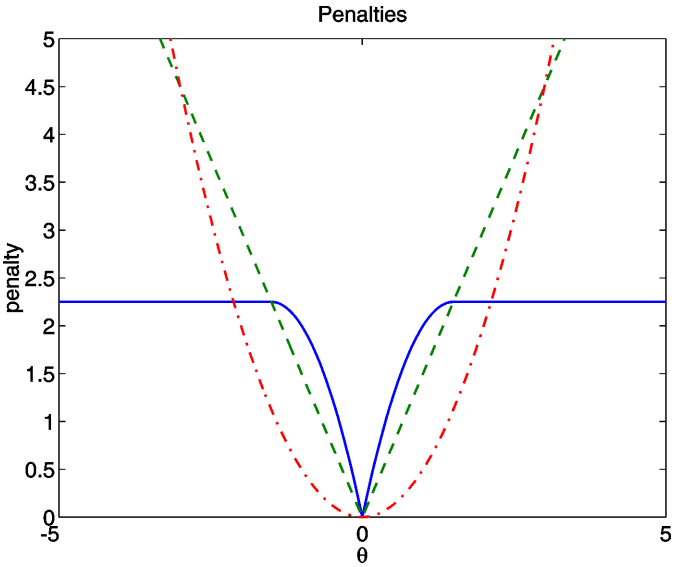

Figure 3. Penalty Functions With λ 1.5, 1, and 1 for the SCAD (——), L_1 (– –), and L_2 (· – ·) Penalties.

3.2 Fisher Scoring Algorithm

Welch et al. (1992) used a stepwise algorithm with the downhill simplex method of maximizing the likelihood function in (4) to sequentially estimate the Gaussian kriging parameters. Here we use a computationally more efficient gradient-based optimization technique to estimate the parameters. The expressions of the gradient and Hessian matrix of the penalized likelihood function in (5) are given in the Appendix. Using the first-order and second-order derivative information, one may directly use the Newton–Raphson algorithm to optimize the penalized likelihood. In this article we use the Fisher scoring algorithm to find the solution of the penalized likelihood because of its simplicity and stability. Notice that $E\{\partial^2\ell(\mu, \boldsymbol{\gamma})/\partial\mu\,\partial\boldsymbol{\gamma}\} = \mathbf{0}$ [see the App. for the expression of $\partial^2\ell(\mu, \boldsymbol{\gamma})/\partial\mu\,\partial\boldsymbol{\gamma}$]. Therefore, the updates of $\hat{\mu}$ and $\hat{\boldsymbol{\gamma}}$ are obtained by solving separate equations. For a given value $(\mu^{(k)}, \hat{\boldsymbol{\theta}}^{(k)}\sigma^{2(k)})$ at the kth step, the new value $(\mu, \boldsymbol{\theta}, \sigma^2)$ is updated by

$$\mu^{(k+1)} = \left\{\mathbf{1}_N^T\mathbf{C}^{-1}\left(\boldsymbol{\theta}^{(k)}\right)\mathbf{1}_N\right\}^{-1}\mathbf{1}_N^T\mathbf{C}^{-1}\left(\boldsymbol{\theta}^{(k)}\right)\mathbf{y} \qquad (17)$$

and

$$\sigma^{2(k+1)} = N^{-1}\left(\mathbf{y} - \mathbf{1}_N\mu^{(k)}\right)^T\mathbf{C}^{-1}(\boldsymbol{\theta})\left(\mathbf{y} - \mathbf{1}_N\mu^{(k)}\right), \qquad (18)$$

where $\mathbf{C}(\boldsymbol{\theta}) = \sigma^{-1}\mathbf{R}(\boldsymbol{\gamma})$, and

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left\{\mathbf{I}_{22}\left(\boldsymbol{\gamma}^{(k)}\right) + \boldsymbol{\Sigma}\left(\boldsymbol{\theta}^{(k)}\right)\right\}^{-1}\partial Q\left(\mu^{(k)}, \boldsymbol{\gamma}^{(k)}\right)/\partial\boldsymbol{\theta},$$

where $\mathbf{I}_{22}(\boldsymbol{\gamma}) = -E\{\partial^2\ell(\mu, \boldsymbol{\theta}, \sigma^2)/\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}\}$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \mathrm{diag}\{p''_\lambda(\theta_1), \ldots, p''_\lambda(\theta_d)\}$, a $d \times d$ diagonal matrix.

3.3 Choice of Regularization Parameter

Because Gaussian kriging gives us an exact fit at the sample point $\mathbf{x}$, the residual at each sample point is exactly equal to 0. Therefore, generalized cross-validation (GCV) cannot be used to choose the regularization parameter λ. In this article we use cross-validation (CV) to select the regularization parameter. We implement V-fold CV, and for a given λ, compute the

V-fold CV score in the following way:

1. Spilt the data $\mathcal{D} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$ 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 the CV score,

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

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

$$\hat{\lambda} = \arg\min_{\lambda_k \in S} CV(\lambda_k).$$

In the literature, leave-one-out CV corresponds to N-fold CV.

3.4 Computing Algorithm for the Proposed Procedure

We summarize the foregoing 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 estimate μ and $\boldsymbol{\gamma}$.
3. Compute the CV score $CV(\lambda_i)$. Let $i = i + 1$.
4. Repeat steps 2 and 3 until all K grid points are exhausted.
5. The final estimator for μ and $\boldsymbol{\gamma}$ is the one with the lowest CV score.

Remark. In step 2, initial values for μ and $\boldsymbol{\gamma}$ are needed. In our implementation, we set $\mu^{(0)} = \bar{y}$ and $\sigma^{2(0)} = N^{-1} \times \sum_{i=1}^{N} (y_i - \bar{y})^2$. Furthermore, we take the initial value for θ_j be θ/σ_j, where σ_j represents the sample standard deviation of the jth component of input vector and θ is the maximizer of $Q(\mu^{(0)}, \theta(\sigma_1, \ldots, \sigma_d), \sigma^{2(0)})$, viewed as a function of θ. Thus θ is easily obtained by plotting θ versus $Q(\mu^{(0)}, \theta(\sigma_1, \ldots, \sigma_d), \sigma^{2(0)})$, because as it is a scalar variable.

4. APPLICATION: PISTON SLAP NOISE

Automobile customer satisfaction is highly dependent on the level of satisfaction that a customer has with the vehicle's engine. The noise, vibration, and harshness (NVH) characteristics of an automobile and its engine are the critical elements of customer dissatisfaction. Piston slap is an unwanted engine noise resulting from 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 piston slap study has shifted from experimental analysis to analytical analysis for both the power cylinder design phase and 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 was given by Hoffman et al. (2003). 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 is modeled using the multibody dynamics code ADAMS/Flex, which also includes a finite-element model. The piston, wrist pin, and connecting rod are 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 is included in the analysis.

We take the piston slap noise to be 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. Because each computer experiment requires intensive computational resources, we used a uniform design (Fang 1980) to construct a design for the computer experiment with 12 runs. We used the centered-L_2 discrepancy uniformity criterion (Fang, Lin, Winker, and Zhang 2000), optimized using the stochastic evolutionary algorithm (Jin, Chen, and Sudjianto 2004). A review of uniform designs and their applications has been given by Fang et al. (2000). The collected data are displayed in Table 1. The ultimate goal of the study is to perform probabilistic design optimization (Hoffman et al. 2003) to desensitize the piston slap noise from the source of variability (e.g., clearance variation), a process that we described in Section 1. To accomplish this goal, the availability of a good meta-model is a necessity. We used a Gaussian kriging model to construct a meta-model as an approximation to the computationally intensive analytical model. In this discussion we focus only on the development of the meta-model. (Interested readers should consult Hoffman et al. 2003 and Du et al. 2004 for the probabilistic design optimization study.) In the data analysis that follows we use $q = 2$, because the response model is smooth.

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	21.0	2	1	.98	56.75
2	15	15.6	21.8	1	2	1.30	57.65
3	29	14.4	25.0	2	1	1.14	53.97
4	85	14.4	21.8	2	3	.66	58.77
5	29	12.0	21.0	3	2	.82	56.34
6	57	12.0	23.4	1	3	.98	56.85
7	85	13.2	24.2	3	2	1.30	56.68
8	71	18.0	25.0	1	2	.82	58.45
9	43	18.0	22.6	3	3	1.14	55.50
10	15	16.8	24.2	2	3	.50	52.77
11	43	13.2	22.6	1	1	.50	57.36
12	57	15.6	23.4	3	1	.66	59.64

4.1 Preliminary Analysis

To quickly gain a rough picture of the logarithm of profile likelihood function (log-likelihood, for short), we set $\theta_j = \theta/\sigma_j$, where σ_j represents the sample standard deviation of the jth component of $\mathbf{x}_i$, $i = 1, \ldots, N$. Such a choice of θ_j allows us to plot the log-likelihood function $\ell(\theta)$ against θ. Plots of the log-likelihood function and the penalized log-likelihood functions with the SCAD, L_1, and L_2 penalties, where $\lambda = .2275$ $[= .5\sqrt{\log(N)/N}\,]$ and $N = 12$, are depicted in Figure 4. The plots suggest that the log-likelihood function reaches its maximum at $\theta = 3$ and is flat when the log-likelihood function near its optimum is flat. In practice, the foregoing approach can be used as a graphical diagnostic tool in determining whether the penalized likelihood approach should be used.

This flat likelihood function creates the same problem exhibited in the simple sinusoidal function example (1) discussed in Section 1 when the sample size equals 6. In contrast, all of the three penalized log-likelihood functions near their optimum are not flat. The resulting penalized MLEs for θ under the constraint $\theta_j = \theta/\sigma_j$ and $\sigma_j \geq 0$ are .0895, .0740, and .0985 for the SCAD, L_1, and L_2 penalties. Although the shapes of the corresponding penalized likelihood functions look very different, their resulting penalized MLEs for θ_j under the constraint $\sigma_j = \theta/\sigma_j$ and $\theta_j \geq 0$ are very close. From the shape of the penalized log-likelihood functions, the resulting estimate of the penalized likelihood with the SCAD penalty may be more efficient than the other two; see Section 2.2 for the relationship between variance and $\boldsymbol{\ell}''(\boldsymbol{\theta})$. This preliminary analysis not only gives us a rough picture of the log-likelihood function and the penalized likelihood function, but also provides us with a good initial value for implementation of the Fisher scoring algorithm. We further demonstrate in next section that the resulting penalized likelihood estimate with the form $\sigma_j = \theta/\sigma_j$ also results in a good prediction rule for the output variable.

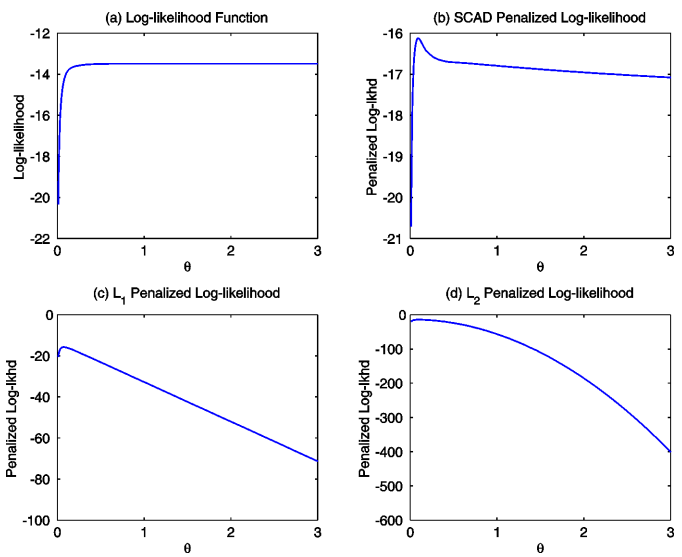

Figure 4. The Log-Likelihood and Penalized Log-Likelihood When $N = 12$. (a) The log-likelihood function. (b), (c), and (d) The penalized log-likelihood functions with the SCAD, L_1, and L_2 penalty.

4.2 Data Analysis via Penalized Gaussian Kriging

We applied the Fisher scoring algorithm with the initial value obtained in the preceding section to the data. We used the leave-one-out CV procedure to estimate the tuning parameter λ. The resulting estimate of λ equals .11, .13, and .06 for the SCAD, L_1, and L_2 penalties. The resulting estimates of μ, σ^2, and θ_j's are given in Table 2. The four estimates for μ are very close, but the four estimates for σ^2 and θ_j's are quite different.

To assess the performance of the penalized Gaussian kriging approach, we conducted another computer experiment with 100 runs. The median of absolute residuals (MAR) is defined as

$$\text{MAR} = \text{median}\big\{|y(\mathbf{x}_i) - \hat{y}(\mathbf{x}_i)| : i = 1, \ldots, 100\big\}. \qquad (19)$$

Equation (19) is used to measure how well the prediction performs. The MAR for the ordinary kriging method is 1.3375, and MAR equals 1.0565, 1.4638, and 1.3114 for the penalized kriging method with the SCAD, L_1, and L_2 penalties. Figure 5 further plots the sorted absolute residuals from penalized kriging versus the absolute residuals from kriging. From these plots, we can see that the penalized kriging with the SCAD uniformly improves the ordinary kriging model. The penalized kriging with the L_2 penalty has almost the same performance as the ordinary kriging model, whereas the penalized kriging with the L_1 penalty does not perform well in this case.

To understand the 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. The plots show that the shape of the log-likelihood function is the same as that of the penalized log-likelihood function with the SCAD penalty, which is in agreement with the theoretical result discussed in Sections 2.1 and 3.1.

We further compute the MLEs for all of the parameters θ_j. Based on five-fold CV, the selected λ equals .18, .105, and .18 for the SCAD, L_1, and L_2 penalties. The resulting estimates, given in Table 3, are all very close, as expected.

4.3 Sensitivity Analysis of Regularization Parameter

Now we examine the sensitivity of the regularization parameter on the prediction results. We concentrate on the penalized kriging method with the SCAD penalty. As reported in Section 4.2, the resulting estimate of λ for SCAD equals .1100. To conduct sensitivity analysis, we reduce and increase the values of regularization parameter by 10%. In other words, we examine the performance of the SCAD penalized kriging with $\lambda = .099$ and $\lambda = .121$. With these two values, we estimate

Table 2. Penalized MLEs

Parameter	*MLE*	*SCAD*	*L_1*	*L_2*
$\hat{\mu}$	56.7275	56.2547	56.5177	56.5321
$\hat{\sigma}^2$	3.4844	4.2345	3.6321	3.4854
$\hat{\theta}_1$	.1397	8.00E–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	3.98E–02	.5779	.2909
$\hat{\theta}_4$	4.0914	5.61E–07	2.02E–04	3.26E–02
$\hat{\theta}_5$	4.0914	3.03E–06	.1501	9.80E–02
$\hat{\theta}_6$	12.2253	4.4979	1.48E–02	.2590

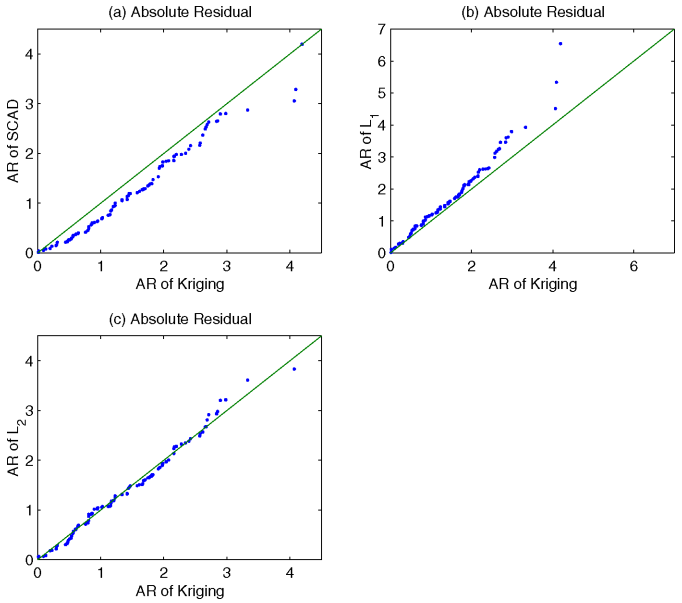

Figure 5. Plots of Absolute Residuals for (a) SCAD, (b) L_1, and (c) L_2.

μ, σ^2, and the θ_j's based on the samples listed in Table 1. The resulting estimates are displayed in Table 4. We also predict the extra 100 response values. The MARs equal 1.0490 for $\lambda = .099$ and 1.0573 for $\lambda = .121$. Note that the MAR for $\lambda = .110$ equals 1.0565. These MAR values are almost similar. The plot of the absolute residuals for these three values of λ, depicted in Figure 7, demonstrate that the absolute residuals are very close for $\lambda = .099$ and $\lambda = .110$ and are almost identical for $\lambda = .110$ and $\lambda = .121$. Figure 7 and the three MAR values indicate that prediction on the extra 100 response values is not sensitive to small changes in the regularization parameter for the SCAD penalized kriging.

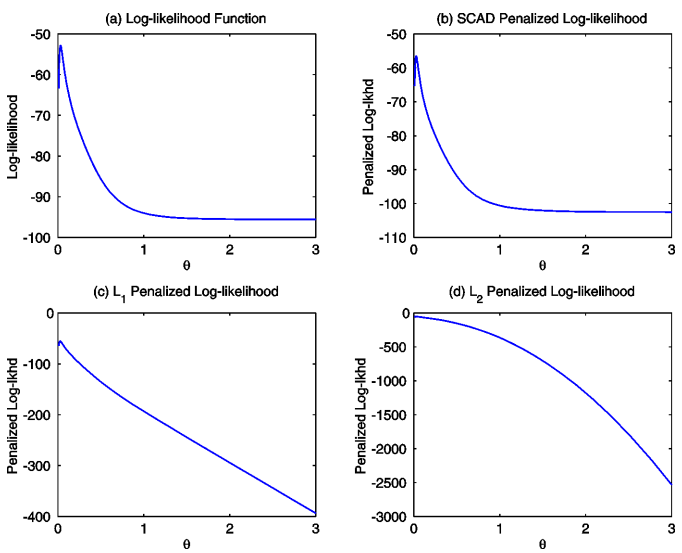

Figure 6. Log-Likelihood and Penalized Log-Likelihood When $N = 100$. (a) The log-likelihood function reaching its maximum at .0314. (b), (c), and (d) Penalized log-likelihood functions with the SCAD, L_1, and L_2 penalty. The locations of the maximum of the penalized likelihood functions are .0314, .0285, and .0314 in (b), (c), and (d).

Table 3. Penalized MLEs

Parameter	*MLE*	*SCAD*	L_1	L_2
$\hat{\theta}_1$	.4514E–4	.3971E–4	.3943E–4	.5858E–4
$\hat{\theta}_2$	.5634E–3	.5192E–3	.5204E–3	.6519E–3
$\hat{\theta}_3$	.3150E–5	.2602E–5	.2618E–5	.4261E–5
$\hat{\theta}_4$	.2880	.2752	.2765	.3003
$\hat{\theta}_5$	.2939E–1	.2593E–1	.2590E–1	.3641E–1
$\hat{\theta}_6$	.1792	.1515	.1548	.2162

5. CONCLUSION

In this article we have proposed SCAD penalized maximum likelihood estimation to deal with problematic flat likelihood functions in Gaussian kriging model parameter estimation. Although REML may be viewed as a kind of penalized likelihood, the motivation of REML is different from that of our penalized likelihood method. For example, the goal of our penalized likelihood method is to reduce the variance of the resulting estimate of θ at the expense of introducing a small bias; however, the goal of REML is to produce an unbiased estimate by paying the price of increased variance of the resulting estimate. This can be easily seen from the REML estimate of the variance of random error (σ^2) for the ordinary multiple linear regression model with independent and identically distributed random errors. We provided practical implementations of the proposed penalized likelihood method, including the Fisher scoring algorithm as well as the choice and sensitivity of the regularization parameter. We also presented comparisons with standard maximum likelihood estimation as well as L_1 and L_2 penalized likelihood using both toy and industrial applications. The method is particularly recommended for constructing a Gaussian kriging metamodel when regular maximum likelihood estimation results are unsatisfactory, a problem commonly encountered when the sample size is small due to computationally expensive engineering simulation models.

ACKNOWLEDGMENTS

The authors thank the editor and the associate editor for their constructive comments and suggestions that have led to significant improvements in the presentation, and to Joseph Stout of Ford Motor Company for permission to use the data in his project. Li's research was supported by National Science Foundation grants DMS-01-02505 and DMS-03-48869. The original manuscript was completed while Sudjianto was working at Ford Motor Company.

Table 4. Penalized MLEs With the SCAD Penalty

Parameter	$\lambda = .099$	$\lambda = .110$	$\lambda = .121$
$\hat{\mu}$	56.2772	56.2547	56.2538
$\hat{\sigma}^2$	4.2745	4.2435	4.2951
$\hat{\theta}_1$	8.28E–04	8.00E–04	7.91E–04
$\hat{\theta}_2$	3.54E–07	1.86E–07	1.70E–07
$\hat{\theta}_3$	3.75E–2	3.98E–02	3.84E–02
$\hat{\theta}_4$	1.02E–7	5.61E–07	7.16E–07
$\hat{\theta}_5$	2.45E–10	3.03E–06	3.71E–08
$\hat{\theta}_6$	4.3917	4.4979	4.4683

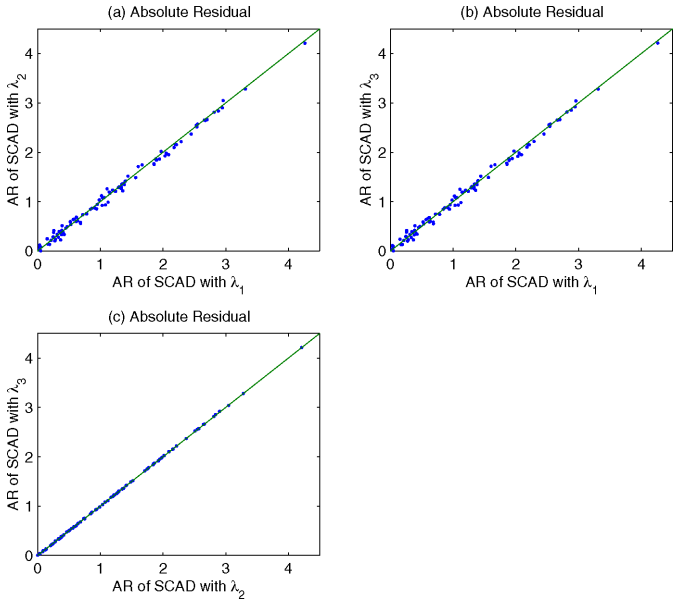

Figure 7. Plot of the Absolute Residuals of the Penalized Kriging With the SCAD Penalty for Three Different Values of λ: $\lambda_1 = .099$, $\lambda_2 = .110$, and $\lambda_3 = .121$. The corresponding estimates of parameters are listed in Table 4.

APPENDIX: DERIVATIVES OF $\ell(\mu, \boldsymbol{\gamma})$

By some straightforward calculations,

$$\frac{\partial \ell(\mu, \boldsymbol{\gamma})}{\partial \mu} = -\mathbf{1}_N^T \mathbf{R}^{-1}(\boldsymbol{\gamma})\mathbf{e}$$

and

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

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

$$\frac{\partial^2 \ell(\mu, \boldsymbol{\gamma})}{\partial \mu^2} = \mathbf{1}_N^T \mathbf{R}^{-1}\mathbf{1}_N;$$

$$\frac{\partial^2 \ell(\mu, \boldsymbol{\gamma})}{\partial \mu\, \partial \gamma_k} = -\mathbf{1}_N^T \mathbf{R}^{-1}(\boldsymbol{\gamma})\dot{\mathbf{R}}_k(\boldsymbol{\gamma})\mathbf{R}^{-1}(\boldsymbol{\gamma})\mathbf{e}, \quad \text{for } k = 1, \ldots, d+1;$$

and

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

where $\ddot{\mathbf{R}}_{ks}(\boldsymbol{\gamma}) = \partial^2 \mathbf{R}(\boldsymbol{\gamma})/\partial\gamma_k\, \partial\gamma_s$.

[Received February 2003. Revised June 2004.]

REFERENCES

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

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, 1–13.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments," *Journal of American Statistical Association*, 86, 953–963.

De Luca, J. C., and Gerges, S. N. Y. (1996), "Piston Slap Excitation: Literature Review," SAE Paper 962396, SAE Transactions.

Du, X., and Chen, W. (2002), "Sequential Optimization and Reliability Assessment Method for Efficient Probabilistic Design," presented at the 2002 ASME Design Automation Conference.

Du, X., Sudjianto, A., and Chen, W. (2004), "An Integrated Framework for Optimization Under Uncertainty Using Inverse Reliability Strategy," *ASME Journal of Mechanical Design*, 126, 1–9.

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

Fang, K. T. (1980), "The Uniform Design: Application of Number-Theoretic Methods in Experimental Design," *Acta Mathematicae Applicatae Sinica*, 3, 363–372.

Fang, K. T., Lin, D. K. J., Winker, P., and Zhang, Y. (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 2003-01-0148, SAE Transactions.

Jin, R., Chen, W., and Simpson, T. W. (2000), "Comparative Studies of Metamodeling Techniques Under Multiple Modeling Criteria," AIAA-2000-4801, presented at the 8th AIAA/NASA/USAF/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Long Beach, CA, September 6–8, 2000.

Jin, R., Chen, W., and Sudjianto, A. (2004), "An Efficient Algorithm for Constructing Optimal Design of Computer Experiments," *Journal of Statistical Planning and Inference*, in press.

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

Kimdeldorf, G. S., and Wahba, G. (1970). "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines," *The Annals of Mathematical Statistics*, 41, 495–502.

Kleijnen, J. P. C. (1987), *Statistical Tools for Simulation Practitioners*, New York: Marcel Decker.

Kodiyalam, S., Yang, R.-J., Gu, L., and Tho, C.-H. (2001), "Large-Scale, Multidisciplinary Optimization of Vehicle System in a Scalable, High-Performance Computing Environment," DETC2001/DAC-21082, 2001 ASME Design Automation Conference, Pittsburgh, PA, September 9–12, 2001.

Koehler, J. R., and Owen, A. B. (1996), "Computer Experiments," in *Handbook of Statistics*, Vol. 13, eds. S. Ghosh and C. R. Rao, Amsterdam: Elsevier Science, pp. 261–308.

Mardia, K. V., and Marshall, R. J. (1984), "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression," *Biometrika*, 71, 135–146.

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

Patterson, H. D., and Thompson, R. (1971), "Recovery of Inter-Block Information When Block Sizes Are Unequal," *Biometrika*, 58, 545–554.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757.

Sacks, 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.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Berlin: Springer-Verlag.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

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

Sweeting, T. J. (1980), "Uniform Asymptotic Normality of the Maximum Likelihood Estimator," *The Annals of Statistics*, 8, 1375–1381.

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

Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B, 40, 364–372.

——— (1990), *Spline Model for Observational Data*, Philadelphia: Society for Industrial & Applied Mathematics.

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, 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, 13–21.

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