E02

Proceedings of the 5th International Conference on Inverse Problems in Engineering: Theory and Practice, Cambridge, UK, 11-15th July 2005

PARAMETER ESTIMATION FOR NOISY DATA AND NUISANCE VARIABLES USING BAYESIAN INFERENCE

A. F. EMERY, E. VALENTI and D. BARDOT
Department of Mechanical Engineering, University of Washington, Seattle, WA 98195-2600, USA
email:emery@u.washington.edu

Abstract - Parameter estimation is generally based upon the maximum likelihood approach and often involves regularization. Typically it is desired that the results be unbiased and of minimum variance. However, it is often better to accept biased estimates that have minimum mean square error. Bayesian inference is an attractive approach that achieves this goal. More importantly, it permits us to consider nuisance variables and incorporates regularization automatically. This paper describes the use of Bayesian inference for an apparently simple experiment that is, in fact, fundamentally difficult and is compounded by a nuisance variable.

1. INTRODUCTION

Let a system, S, have a measurable response, R, that is $R = S(\mathbf{x}, t, p)$ where $\mathbf{x}$ denotes spatial position, t denotes time, and p denotes parameters. A model M is to be constructed that is presumed to accurately reflect the systems behavior such that $R = M \equiv S(\mathbf{x}, t, P, \Theta, \mathcal{N})$ where P and Θ represent parameters that are known and to be estimated, respectively, and $\mathcal{N}$ represents *nuisance* parameters that affect the model but which we are not interested in estimating. Obviously we could estimate both Θ and $\mathcal{N}$, but this would generally require more data and for many cases create additional instabilities in the solution. The classification of a parameter as belonging to P, Θ, or to $\mathcal{N}$ depends upon the specific situation.

The sensitivity of the response to the parameters can be best characterized by the relationship between their uncertainties through the usual equation for variance, written here for two parameters, θ_1 and θ_2 with standard deviations of $\sigma(\theta_1)$ and $\sigma(\theta_2)$, as

$$\sigma^2(R) = (\frac{\partial R}{\partial \theta_1})^2 \sigma^2(\theta_1) + (\frac{\partial R}{\partial \theta_2})^2 \sigma^2(\theta_2) + (\frac{\partial R}{\partial \theta_1})(\frac{\partial R}{\partial \theta_2}) cov(\theta_1, \theta_2) \tag{1}$$

If the measured response is noisy, i.e. $\sigma(R) > 0$, then the uncertainty in the parameters, $\sigma(\theta)$, is inversely proportional to the sensitivity $\partial R/\partial \theta$. While it is axiomatic that parameters should be estimated only for cases in which the sensitivity is high, there are many cases in which the sensitivities are unavoidably small, the resulting uncertainty in the parameters is large, and the stability of the estimation methods is poor. Most methods involve some form of regularization to stabilize the solution and then appeal to conventional statistics to define the resulting estimates in terms of point values, i.e., average values and confidence limits [1]. For many problems this approach suffices, particularly when based upon Gaussian statistics, but in some situations the statistical characterization is insufficient to accurately define the uncertainty of the estimates.

In this paper we advocate the use of Bayesian inference, often referred to as Stochastic (or Statistical) Regularization [2, 3], for estimating parameters, especially when nuisance variables have to be considered. Bayesian inference has a long and well developed history, mostly mired in controversy because of its "subjective" character and the need to utilize simplistic and often limiting probability distributions due to the difficulty in evaluating the resulting inferences. This was particularly true for complex models. However, with the computing power now available, one can numerically evaluate many of the complex features of the method and obtain the desired results.

We will see that estimates derived from Bayesian inference have minimum mean square error, are regularized, and account for nuisance variables in a consistent and theoretically sound manner. While the method involves additional computations, the dramatic increase in computing power makes robust Bayesian inference not only possible, but to be encouraged.

1.1 Test Case–The Error in Variables Model (EVM)

To demonstrate the use of Bayesian inference we have chosen the simple and apparently *innocuous* problem of estimating the thermal conductivity by measuring the heat flux through and the temperature difference across a layer of cross-sectional area A in a one dimensional conduction test system, Figure 1. The conductivity is found by

$$k = \frac{(Q - Q_L)t/A}{\Delta T} \text{ which we write as } = \frac{q(1-f)}{\Delta T}, \text{ where } q \equiv \frac{Qt}{A} \tag{2}$$

where Q_L represents the heat lost. Since the heat lost is proportional to the mean temperature of the layer, it is also proportional to the heat input and we take it as $Q_L = fQ$. The constant of proportionality, f, is, in general unknown, since it depends on the specifics of the ambient conditions (convective and radiative losses) and how the layer is embedded in the 'guarded heat box.' Thus f is a nuisance variable and all that we probably know about it is some 'quoted' average value, that f is always positive and less than one, and probably not more than a few multiples of the quoted value. Our model of the system in terms of measurable variables is given by

$$\Delta T = \frac{q(1-f)}{k} \equiv \beta q(1-f) \text{ where } \beta \equiv \frac{1}{k} \tag{3}$$

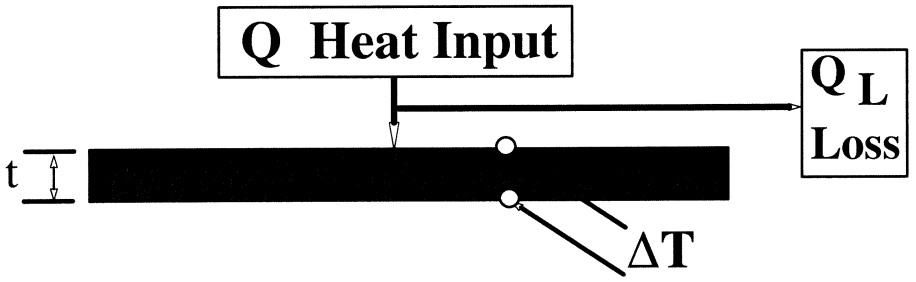

Figure 1: Schematic of conductivity measuring experiment.

In reality, neither q nor ΔT can be measured without error. Thus, our system is really defined in terms of the measured responses, R^q and R^T, by

$$R_i^q = q_i + \epsilon_i^q \tag{4a}$$
$$R_i^T = \Delta T_i + \epsilon_i^T \tag{4b}$$

Note that f depends upon the conditions of the experiment and might be regarded as a stochastic quantity. However, for a given set of test conditions we are going to treat it as a fixed, but unknown value, i.e., $\sigma(f) = 0$. Assuming the errors, ϵ^q and ϵ^T, to be of zero mean, constant variance, and uncorrelated, our first guess of the imprecision of our estimate of k, denoted by $\hat{k}$, using eqn.(1) would be given by

$$\sigma^2(\hat{k}) = \left(\frac{\partial k}{\partial q}\right)^2 \sigma^2(q) + \left(\frac{\partial k}{\partial \Delta T}\right)^2 \sigma^2(\Delta T) \tag{5}$$

For an aluminum alloy 2024-T6, with a conductivity of 186 W/mK [4], a 2 cm layer with $q = 1200 W/mK$ gives a $\Delta T \approx 6C$. Assuming that the heat input, q, can be measured with an accuracy of 2.5% and that the ΔT can be measured to within 0.5 °C, [5], eqn.(5) gives $\sigma(\hat{k})/\hat{k} \approx 8\%$. Now if N measurements are taken, the standard deviation of the average $\hat{k}$ equals $\sigma(\hat{k})/\sqrt{N}$ and for N=21, this would give $\sigma(\hat{k})/\hat{k} \approx$ 1.8%. Taking the usual 95% confidence interval based on the student-t distribution for 20 degrees of freedom, $\pm 2.08\sigma(\hat{k})/\hat{k} \approx 4\%$ or a range of $179 \leq \hat{k} \leq 193$. We will see that we can do better with Bayesian inference.

1.2 Error in Variables Models - EVM

Now the experiment can be run in two different ways: Procedure A, and Procedure B. In Procedure A, N repeated independent experiments are run with the heat input set to a nominally constant value. Even though q_i is presumed to be constant it varies randomly about a mean value, for example because of fluctuations in the power supply (e.g. varying line voltages), and is measured with error.

Thus both R^q and R^T are realizations of random processes about a constant mean and with a constant standard deviation. Figure 2a illustrates a sample data set and the least squares fit of eqn.(3). The computed standard deviation is 1.8% and the confidence interval is $176 \leq k \leq 195$. Note that it is critical that the correct model be used. If one uses a linear regression that is not forced through the origin, as required by eqn.(3), the estimate of k is clearly wrong.

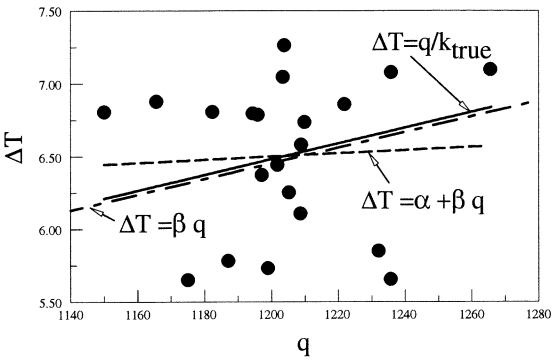

Figure 2a: Data and fit for procedure A.

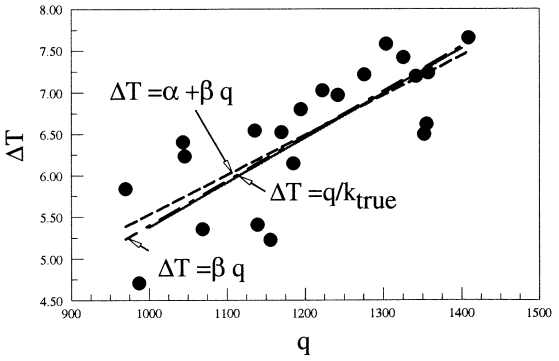

Figure 2b: Data and fit for procedure B.

In Procedure B, N tests are run with different values of heat input, which are assumed to be constant during each test. These values of q_i are deterministic but known only approximately because of the measurement errors. A sample set of data is illustrated in Figure 2b. Using eqn.(3) in the least squares fit, the results are the same as for Procedure A. In this procedure the range of q_i is sufficient such that forcing the fit through the origin is not critical in estimating k, but not forcing it to do so will lead to a standard deviation approximately 10 times larger than that given by fitting eqn.(3).

Since $Q - Q_L$ and ΔT are related through eqn.(3), estimates of $\sigma(\hat{k})$ based upon eqn.(2) are incorrect because of the neglect of the correlation between R^q and R^T. Equations (4) can be written as

$$R_i^q = q_i + \epsilon_i^q \tag{6a}$$

$$R_i^T = \beta q_i + \epsilon_i^T = \beta R_i^q - \beta \epsilon_i^q + \epsilon_i^T \tag{6b}$$

Equations (6) constitute what is referred to as the 'error in variables model' that has been extensively treated from both the classical [6] and Bayesian [7] points of view. These models are generally classified as: 1) *structural* where q_i are random variables, independent of the errors, having a constant mean and a constant variance of σ_s^2, i.e., Procedure A*, 2) *functional* where q_i represent unknown constants, Procedure B. From the classical point of view, the models must be treated differently in estimating β [6] and require additional information about either $\sigma(\epsilon^q)$, $\sigma(\epsilon^T)$, or their ratio, none of which are likely to be known with exactitude and generally are estimated from the measurements.

The true values of q_i and ΔT_i are not known because of the measurement errors. However, because of the model, eqn.(3), that relates q_i and ΔT_i, the measurements, R_i^q and R_i^T, are correlated with a covariance of $-\beta\ \sigma_q^2$. Including this term in eqn.(1), gives a reduction of approximately 40% in the standard deviation as shown in Table 1.

Table 1. Estimated uncertainties in conductivities, eqn.(1). True value = 186.

	$\hat{k}$	$\sigma(\hat{k})$	Confidence Interval
eqn.(1) w/o cov		3.26	179-193
eqn.(1) w/ cov		2.10	182-190
Least Squares Fit	185.3	3.33	176-195

These values were obtained because we knew the standard deviations of the measurement errors. The covariance term is important because of its large negative value relative to the other terms. If $\sigma(\epsilon^q)$ and $\sigma(\epsilon^T)$, and k are poorly known, it is not possible to estimate the covariance term $-\beta\sigma_q^2$. The results given in the table are for no losses, i.e., $f = 0$. The question is how to account for the uncertainty in the losses.

1.3 Conventional Estimation

Although the relationship between regularization and Bayesian inference has been noted before, it is useful to review the usual methods of parameter estimation based upon the least squares approach to emphasize the relationship of Bayesian Inference to regularization. Consider the case where the model M is a non-linear function of the parameters Θ. Let the true value of the parameter be denoted by Θ and the estimated value by $\hat{\Theta}$. The measurements are presumed to be corrupted by the noise ϵ to give

* From eqn.(5a), the variance of R^q is $\sigma_s^2 + \sigma_q^2$

$$R = M(\Theta) + \epsilon \tag{7}$$

where $E[\epsilon] = 0$ and $cov[\epsilon] = \Sigma$. The estimated values, $\hat{\Theta}$ are those that minimize the functional, $F(\Theta)$

$$F(\hat{\Theta}) \equiv ||R - M(\hat{\Theta})|| \tag{8a}$$

$$= ||R - M(\Theta) - \frac{dM}{d\Theta}|_{\Theta}(\hat{\Theta} - \Theta)|| \tag{8b}$$

where $||z||$ represents the length of the vector z and

$$R - M(\hat{\Theta}) = e, \quad \text{the residual} \tag{8c}$$

Within the linear assumption, provided that the number of readings, N, is sufficiently large [8] it is permissible to evaluate $A_j = dM/d\Theta$ at $\Theta = \hat{\Theta}_j$, and the equation is solved iteratively using*

$$\hat{\Theta}_{j+1} = \hat{\Theta}_j - (A_j^T \Sigma^{-1} A_j)^{-1} A_j^T \Sigma^{-1} [R - M(\hat{\Theta}_j)] \tag{9}$$

For d parameters, R and M are [Nx1] vectors, Θ is a [1xd] row vector and A is a [Nxd] matrix. Upon convergence, the estimate $\hat{\Theta}_b$ satisfies

$$\hat{\Theta}_b - \Theta = (A_j^T \Sigma^{-1} A)^{-1} A^T \Sigma^{-1} [R - M(\hat{\Theta}_b)] \tag{10a}$$

$$E[\hat{\Theta}_b] = \Theta \tag{10b}$$

$$cov[\hat{\Theta}_b] = (A^T \Sigma^{-1} A)^{-1} \tag{10c}$$

Equations (10b) and (10c) represent the generally acceptable desirable state of $\hat{\Theta}$ being an unbiased estimator with minimum variance – i.e., satisfying the Cramer-Rao lower bound [9]. Such estimators are termed "efficient" and are the most informative, where the Fisher information matrix, I, is defined as $I = cov^{-1}(\Theta) = A^T \Sigma^{-1} A$. These estimators are generally called BLUE, Best Linear Unbiased Estimators, hence the notation $\hat{\Theta}_b$.

In solving eqn.(9) one must know the covariance matrix Σ. Generally, the best that can be done is to write

$$\Sigma = \sigma_n^2 \Omega \tag{11}$$

where Ω is the correlation matrix and σ_n is the standard deviation of the noise. Substituting for Σ from eqn.(11) into eqn.(9), σ_n^2 cancels and only the correlation matrix is needed to obtain $\hat{\Theta}$. If all the noise is from a single source, the correlation matrix is often taken to be the identity matrix, I. Although $\hat{\Theta}$ can be found without knowing σ_n, it is required to determine the variance of $\hat{\Theta}$. If not given, an approximate value can be found from the residuals by using

$$\hat{\sigma}_n^2 = \frac{e^T \Omega^{-1} e}{N - d} \tag{12}$$

where the superscript T denotes transpose and d denotes the number of parameters estimated. To arrive at the results, eqns (10) and (12), the only statistical information needed is the correlation matrix of the measurement noise. This can usually be obtained by standard statistical analyses [10]. Although most reported parameter estimation studies that considered real data ignored possible correlations, it is not uncommon for sequential measurements taken at a relatively high sampling rate to be correlated. The effect of correlated measurements is to increase the standard deviation of the estimated parameter. Let the measurement errors be autoregressive with $cov(R_i, R_j) = \rho^{|i-j|}$, not uncommon for reasonably high sampling rates. Emery [11] examined a sequence of transient temperature measurements and showed that a considerable correlation, $\rho > 0.2$, existed. The effect of the correlation is to increase the standard deviation of $\hat{\Theta}$, giving

$$\sigma^2_{with\ correlation} = \sigma_n^2 (\frac{1+\rho}{1-\rho}) \tag{13}$$

When the readings are correlated, one can look at the problem as one with an increased σ_n of the data or as one in which the effective number of readings has been diminished to $N_{effective} = N(1-\rho)/(1+\rho)$.

At this point, all that one can say about the estimate, $\hat{\Theta}$, is its value and an approximate value of the standard deviation. Most investigators will then cite the usual equations for confidence limits based on the student-t distribution to give some idea of the range of $\hat{\Theta}$ [9].

---

* Because of the reparameterization from k to β, eqn.(3) is linear in β and no iteration is needed

2. MAXIMUM LIKELIHOOD

The results given in eqn.(10) are based upon classical statistics and they reflect what one would expect to find if an experiment were conducted a great number of times, with each time producing an estimate $\hat{\Theta}_i$. Most experiments are conducted only a few times and the concept of an expected value, $E[\hat{\Theta}]$, is not appropriate. In the Maximum Likelihood method the estimated parameter is taken to be that value for which the data actually measured had the highest probability of occurring. That is, given a conditional probability distribution of the data $l \equiv p(R|\Theta)$*, one searches for values of Θ that maximize l. Since the data, R, are composed of a deterministic part, $M(\Theta)$, and a random error, ϵ, the probability density distribution of R is simply that of ϵ. We must then postulate a pdf for the error. The most common assumption is that the errors are independent and identically distributed and follow a Gaussian distribution with zero mean and a constant standard deviation, but may be correlated. This leads to a likelihood of

$$p(R|\Theta) = \frac{1}{\sqrt{2\pi}^N \sqrt{det(\Sigma)}} \, e^{-\frac{1}{2}(R-M(\Theta))^T \Sigma^{-1} (R-M(\Theta))} \tag{14}$$

The values of $\hat{\Theta}_{MLE}$ maximize eqn.(14). The maximum likelihood method is probably the single most used method in parameter estimation. In principle, it can be used to estimate both Θ and Σ and any other unknown parameters in the model. If Σ is known, $\hat{\Theta}_{MLE}$ are unbiased. However, estimates of the components of Σ are not unbiased, except in the limit as $N \rightarrow \infty$, i.e., asymptotically unbiased [8, 12].

For Procedure A, maximum likelihood can be used to estimate $\beta, \sigma_T, \sigma_q, q,$ and ΔT only if additional information is available. For the case of eqn.(3), forcing the intercept to be zero is sufficient information for estimating β. For Procedure B, the functional model, we are attempting to estimate $\beta, \sigma_T, \sigma_q$ and the N values of q_i. It has been shown that the likelihood has no maximum, only a saddle point [6, 12] and thus Θ cannot be estimated. Since every new test introduces another q_i, more data will not resolve the issue. However if the ratio, σ_T/σ_q is known, a solution is possible [6]. This is also true if the structural model includes more than one parameter.

Nuisance variables *cannot* be treated by maximum likelihood. About the best that can be done is to first estimate all parameters, substitute $\hat{\mathcal{N}}$ into eqn.(14), and then re-estimate $\hat{\Theta}$ [12].

3. TIKHONOV REGULARIZATION

If the matrix $A^T\Sigma^{-1}A$ is ill-conditioned, one sees from eqn.(10c) that the standard deviation of $\hat{\Theta}$ is likely to be very large. The usual approach is to regularize the equation by minimizing

$$F(\Theta) = ||R - M(\hat{\Theta})|| + \alpha \mathcal{S} \tag{15}$$

where $\mathcal{S}$ is a stabilizing functional and α is a problem dependent constant. The choice of $\mathcal{S}$ depends upon qualitative assumptions about Θ and strongly influences both the solution and its convergence [13]. $\mathcal{S}$ is subject to some fairly restrictive conditions to ensure that a solution can be obtained [14]. For many engineering problems, $\mathcal{S}$ is often taken to be a sum of zeroth, first, and second order derivatives of $\hat{\Theta}$ which can be represented in the form

$$\mathcal{S}(\Theta) = ||(\Theta - \overline{\Theta})^T \Phi (\Theta - \overline{\Theta})|| \tag{16}$$

where $\overline{\Theta}$ is a chosen value, Φ is a symmetric positive definite matrix satisfying $\Theta^T \Phi \Theta \geq \gamma ||\Theta||^2$ for a fixed γ and all Θ. In this case Φ is a diagonal matrix with 3, 5, and 7 diagonals respectively [14]. The resulting equation to estimate Θ is then

$$\hat{\Theta}_t - \Theta = (A^T\Sigma^{-1}A + \alpha \Phi^T \Phi)^{-1} A^T \Sigma^{-1} (R - M(\hat{\Theta}_t)) \tag{17}$$

In essence, one is imposing a requirement that Θ be a smooth function. We shall see that the form of eqn.(17) corresponds directly to the assumption of prior information used in Bayesian inference.

The estimate, $\hat{\Theta}_t$, is biased since the norm $||\hat{\Theta} - \Theta|| \leq \delta/\sqrt{\alpha}$, where δ is an error bound on $||\epsilon||$. This illustrates the dilemma: choosing a small α leads to instability, but a large value overly smoothes the solution. A solution in which $\alpha = \alpha(\delta)$ is termed a regular algorithm. It is common to use Morozov's discrepancy principle [13, 14] which suggests that α be chosen such that

$$||R - M(\hat{\Theta})|| \approx \delta \tag{18}$$

Invoking Morozov's discrepancy principle leads to $||\hat{\Theta} - \Theta|| = O(\sqrt{\delta})$. Other than this general result, it is difficult to understand how α affects the results, except by examining specific problems. Since Tikhonov

* $p(R|\Theta)$ denotes the pdf of R given a specified value of Θ

regularization can be regarded as a form of ridge regression, it is useful to appeal to this method for insight into regularization. Ordinary ridge regression [15] yields the biased estimator $\hat{\Theta}_r$ where

$$\hat{\Theta}_r - \Theta = (A^T\Sigma^{-1}A + \rho I)^{-1}A^T\Sigma^{-1}(R - M(\hat{\Theta}_r)) \tag{19}$$

If the errors are uncorrelated and of constant variance, $\Sigma = \sigma^2 I$, a comparison of eqns (17) and (19) shows that if $\Phi = I$, $\rho = \alpha$. The importance of this observation is that an appropriate choice of ρ gives a parameter with a minimum mean square error that is less than that of $\hat{\Theta}_b$

$$E[||\hat{\Theta}_r - \Theta||] = var(\hat{\Theta}) + bias^2(\Theta) \leq E[||\hat{\Theta}_b - \Theta||] \tag{20}$$

It is common to use values of ρ between 0.7 and 0.95 σ^2. While it is always possible to find a value of ρ, i.e, α, that will minimize $E[||\hat{\Theta}_r - \Theta||]$ for a specific Θ, it is not possible to find one value that will suffice for all Θ for a given model, $M(\Theta)$ [14].

4. BAYESIAN INFERENCE

Bayesian inference is based on Bayes' equation relating conditional probabilities, illustrated here for two independent parameters, θ_1 and θ_2 and a response R

$$p(\theta_1, \theta_2|R) = \frac{p(R|\theta_1, \theta_2)\pi(\theta_1)\pi(\theta_2)}{C} \qquad \text{where } C = \int\int p(R|\theta_1, \theta_2)\pi(\theta_1)\pi(\theta_2)d\theta_1 d\theta_2 \tag{21}$$

$\pi(\theta_1)$ and $\pi(\theta_2)$ are termed the 'prior' probabilities and reflect information about θ before the experiment is run. $p(\theta_1, \theta_2|R)$ is called the 'posterior' probability of θ_1 and θ_2 and reflects how the priors are modified by the experimental data. If the errors are assumed to be normally distributed and a prior which assumes that Θ is normally distributed about a mean of Γ with a variance of V is used, the posterior is

$$p(\Theta|R) = \frac{1}{C}\frac{1}{(\sqrt{2\pi})^N\sqrt{det(\Sigma)}}e^{-\frac{1}{2}(R-M(\Theta))^T\Sigma^{-1}(R-M(\Theta))}\frac{1}{\sqrt{2\pi}^d\sqrt{det(V)}}e^{-\frac{1}{2}(\Theta-\Gamma)^TV^{-1}(\Theta-\Gamma)} \tag{22}$$

Upon evaluating C, we find that $\hat{\Theta} - \Gamma$ is normally distributed, leading to the Bayes' estimator

$$\hat{\Theta}_B = \Gamma + (V^{-1} + A^T\Sigma^{-1}A)^{-1}A^T\Sigma^{-1}(R - M(\hat{\Theta}_B)) \tag{23a}$$

with a variance of

$$var(\hat{\Theta}_B) = (V^{-1} + A^T\Sigma^{-1}A)^{-1} \tag{23b}$$

Comparing eqns (17) and (23) reveals the direct connection between Bayesian inference and the traditional regularization approach. Bayesian inference contains all of the needed requirements, regularization, and the possibility of minimum mean square error, with V being the direct equivalent of $\alpha\Phi^T\Phi$ in eqn.(17).

A particular problem is the evaluation of C in eqn.(14) for complex nonlinear models, $M(\Theta)$. If one were content with characterizing Θ by the mode of the posterior, then it is only necessary to find the maximum of the numerator and C is unimportant. It is generally accepted that a better representation of $\hat{\Theta}$ is the mean, i.e., $\hat{\Theta} = \int p(\Theta|R)d\Theta$, whose evaluation requires evaluating C. Furthermore, specifying either the mode or the mean without an estimate of the confidence limits is unsatisfactory. Bayesians refer to these limits as high posterior density (HPD) limits or credible limits. To find these limits, as depicted on Figure 3, it is necessary to know $p(\Theta|R)$ in order to evaluate the integral. Typically Monte Carlo methods are used, but they are computationally expensive. The Markov Chain Monte Carlo approach [16] has provided some relief in this area as demonstrated in [17].

Probably the most subjective part of Bayesian inference is the choice of the priors. Consider the case of estimating the mean of a number of readings. If the 'noninformative' priors [18, 19] are used (i.e., a constant for the mean and $1/\sigma$ for the standard deviation) the estimate is identical to the classical value. However this correspondence between Bayesian and classical estimators of other statistics is not true in general. When 'proper' priors, i.e. ones that integrate to unity, are used, Wald [20] has shown that the estimator has the smallest mean square error of any estimator over the same range of variables.

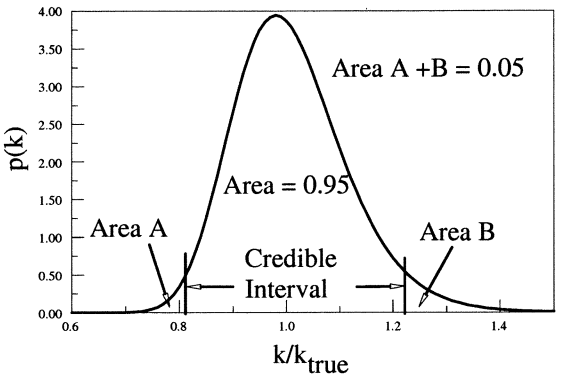

Figure 3: Illustrating the 95% high posterior density region (Credible Interval) for $p(k)$ where $k = Q/\Delta T$ and with Q and ΔT normally distributed.

4.1 Nuisance Parameters

A very important feature of Bayesian inference is the treatment of 'nuisance parameters'. These are parameters that affect the posterior but are not of interest, i.e. are nuisance variables. For example if the standard deviation of the errors, σ, is not known and is of no interest, the posterior is 'marginalized' through integration. Letting θ_2 be σ, then the marginal posterior pdf of θ_1 is

$$p(\theta_1|R) = \int p(\theta_1, \sigma|R)d\sigma = \int p(\theta_1|R, \sigma)\pi(\sigma)d\sigma \tag{24}$$

In solving eqn.(24), a prior for the nuisance variable must be specified. Priors range from the noninformative to highly specific pdfs.

4.2 Optimal Design

Optimal design refers to the choice of the measurement points, $\mathbf{x}, t$ and known parameters, P, that will minimize $var(\Theta)$. For a single parameter, $A^T\Sigma^{-1}A$ is a scalar, but even then finding the optimal values of $\mathbf{x}, t$ is not a trivial task since the contribution of different measurement points, $\mathbf{x}$, will change with time. Thus, other than gross assumptions about t_i, it is rarely possible to define an optimal design. For d parameters, $W = A^T\Sigma^{-1}A$ is a matrix and two common metrics for multi-parameter inverse design are the A and D optimalities [21]. D minimizes the determinant of W, i.e., the product of the eigenvalues. If any eigenvalue of W is zero, D will be zero regardless of the uncertainty of the other parameters. A optimality minimizes the trace of W, i.e., the sum of the eigenvalues and can only equal zero if all are zero. From the classical point of view it is hard to predict the effect of the measurement points on these different measures.

In Bayesian inference, optimal design is defined by minimizing $W^* = (V^{-1} + A^T\Sigma^{-1}A)$. Now it is easy to see that if good information is posited about any one parameter, θ_d, that the diagonal element $V(d,d) \rightarrow 0$ and the determinant of W^* will approach zero regardless of the imprecision in the other parameters. However, if A optimality is used, it can only approach zero if all variances go to zero, meaning perfect prior knowledge of all parameters. A optimality is not without its problems since it is not invariant to reparameterization [18]. Thus an optimal design for β, eqn.(3), may not be optimal for k. However, recognition that the optimality is related to the prior information, and thus in principle to the design points, is often useful in choosing the points [22].

A good example of combining optimal design with Bayesian techniques is given by Sacks [22] where Gaussian processes were used for the design of a complicated electrical circuit. The optimal points were found by an exchange process in which a first choice of design points was made. Then each design point was varied slightly and the optimality computed in terms of the variance. If the variance was reduced, the new design point was accepted. This was continued until no better design point could be found. Attention was then directed to another design point and the process repeated. The process is not simple, usually is very time consuming, and does not yield a unique set of design points.

For statistical problems, the concept of optimal design points is very relevant. For engineering and scientific experiments, the range of specific parameters is usually very restricted and finding optimal points is not always relevant. This is particularly true when responses are measured with respect to position and time. Sampling times are almost always in uniform increments and measurement points are generally fixed. In the problem discussed here, where it is possible to explore the effect of different applied heating rates, we have chosen to use regular increments for simplicity.

5. CASE STUDY – ESTIMATION OF k

Since the conventional least squares approach (fitting ΔT_i to q_i) assumes that q_i is without error, let us look at the problem from the Bayesian point of view. Assuming that the errors are normally distributed, the likelihood is

$$p(R^T, R^q|q, \beta, f, \sigma_q, \sigma_T) \propto \frac{1}{\sigma_q^N}\frac{1}{\sigma_T^N} e^{-\frac{1}{2\sigma_q^2}(R^q-q)^T(R^q-q)} e^{-\frac{1}{2\sigma_T^2}(R^T-\Delta T)^T(R^T-\Delta T)} \tag{25}$$

While one may have a good idea of σ_q, σ_T is more problematic. From a knowledge of the instrumentation one often has a reasonable idea of $\sigma_q/\ \sigma_T$. Let us recast the problem in terms of $\phi = \sigma_q^2/\sigma_T^2$ giving a posterior of

$$p(q, \beta, f, \phi, \sigma_q|R^T, R^q) = \frac{p(R^T, R^q|q, \beta, f, \phi, \sigma_q)\pi(q)\pi(\beta)\pi(f)\pi(\phi)\pi(\sigma_q)}{\int p(R^T, R^q|q, \beta, f, \phi, \sigma_q)\pi(q)\pi(\beta)\pi(f)\pi(\phi)\pi(\sigma_q) dq\ d\beta\ df\ d\phi\ d\sigma_q} \tag{26}$$

We marginalize to get the posterior $p(\beta|R)$ by integrating over q, ϕ, σ_q, and f: a non-trivial task which can usually be done only numerically. As noted by Zellner [7], the posterior $p(\beta)$ may depend strongly upon the prior for q. In a rough sense, the extremes are a) considering $\pi(q)$ as uniform over $-\infty \leq q \leq \infty$ (an improper pdf, although $p(\beta)$ will be proper) and b) concentrated about the maximum likelihood estimate of q, which for a given value of ϕ is

$$\hat{q}_{MLE} = \frac{\phi R^q + \beta\ R^T}{\phi + \beta^2} \tag{27}$$

Using non-informative priors for β, f, q, ϕ and σ_q and integrating gives

$$p(\beta, f, \phi|R) \propto \frac{1}{\phi}[(R^T - \beta(1-f)R^q)^T(R^T - \beta(1-f)R^q))]^{-N/2} \tag{28}$$

If the more concentrated assumption is used, $q = \hat{q}_{MLE}$, the corresponding posterior is

$$p(\beta, f, \phi|R) \propto \frac{\phi^{N/2}(1+\beta^2(1+f)^2/\phi)^N}{[(R^T - \beta(1-f)R^q)^T(R^T - \beta(1-f)R^q)]^N} \tag{29}$$

It is clear that our estimates of $\hat{\beta}$ will depend upon what priors we assume. The robustness of the approach was tested by using uniform and normal priors for β, f, ϕ and also an inverted Gamma distribution for ϕ. For *f=0*, no losses, the posterior $p(\hat{k}|R)$ was found to be essentially normal as shown in Figure 4, but with a narrow confidence interval. Estimates of $\hat{k}$ were quite insensitive to the form of the priors and some representative values are shown in Table 2. It is because of the narrowness of $p(\hat{k}|R)$ that the confidence intervals listed in Table 2 are much tighter than those from least square regression.

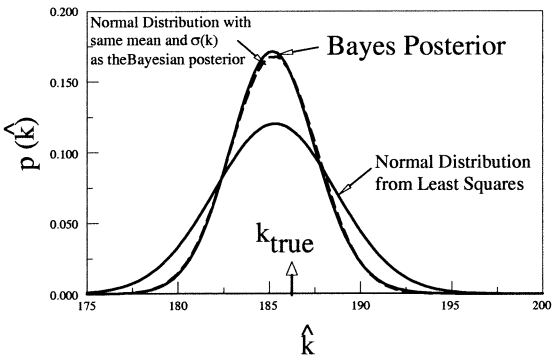

Figure 4: The posterior $p(\hat{k})$ with no losses.

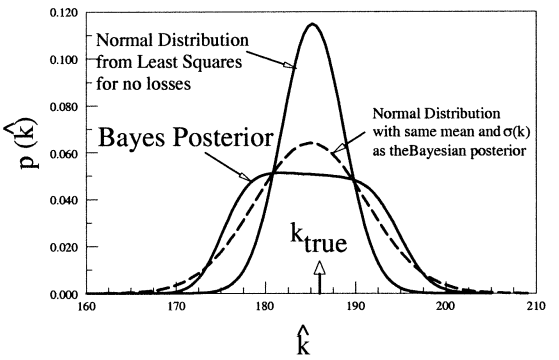

Figure 5: The posterior $p(\hat{k})$ with losses, $0 \leq f \leq 0.1$.

When losses are present we assumed that the quoted heat loss for the system was an average of 5%. For high conductivity values it should approach 0 and for insulators with large ΔT we assumed that it could approach 10%. Using a uniform prior from 0 to 10%, an inverted Gamma, and a normal distribution gave very similar results. Figure 5 illustrates the posterior pdf of $\hat{k}$. The posterior $p(\hat{k}|R)$ is far from normal and, as expected, the uncertainty in f leads to an increased standard deviation of $\hat{k}$ and a slight bias in $\hat{k}$. For no heat loss, or equivalently a known heat loss, the estimated conductivity was in error by less than 1% and when the heat loss was uncertain, in error by less than 2%. Because of the shape of the Bayesian posterior, there is not much difference between the 95% and 99% confidence intervals, in contrast to the width of the confidence interval for a normal distribution.

Table 2. Estimated conductivities, True value = 186.

Type	$\pi(q_i)$	Mean Value	$\sigma(\hat{k})$	95% Confidence Interval	99% Confidence Interval
		No Losses			
Structural	Uniform	184.9	2.38	180-190	179-191
" "	Least Squares Fit	185.1	3.33	178-192	172-198
Functional	Uniform	185.2	2.37	181-190	179-192
" "	$q = q_{MLE}$	185.2	3.45	179-192	177-195
" "	$N(R^Q, 4\sigma_q^2)$	185.2	2.37	181-190	179-192
" "	Least Squares Fit	185.6	3.33	176-195	173-198
		Losses, $0 \leq f \leq 10\%$			
Structural	Uniform	181.7	6.22	174-196	172-199
Functional	Uniform	181.7	6.22	174-196	172-199
" "	$q = q_{MLE}$	183.8	6.76	173-198	172-201
" "	$N(R^Q, 4\sigma_q^2)$	181.5	6.22	174-196	172-199

where $N(a, b)$ denotes a normal distribution with a mean of a and a standard deviation of b

6. CONCLUSIONS

As tolerances become tighter and risk avoidance is emphasized, designs are increasingly focusing on estimating sensitivities to the many parameters of the model. To date most analyses have used the uncertainty propagation, eqn.(1), based upon some nominal values of the parameters. Unfortunately this gives only local sensitivity and may not reflect the overall behavior. Such behavior can only be found from the posterior pdf of the system's responses as determined from eqn.(28). This requires a realistic prior. In many situations, the assumption that the posterior, $p(k|R)$ is normal is reasonable as shown on Figure 4 when there is no uncertainty about the losses.* However, when some parameters have posterior distributions that are non-normal, Figure 5, conclusions about system sensitivity may be seriously in error.

Bayesian inference as commonly applied in statistical regularization, has the ability to automatically include regularization and to yield a minimum mean square error. Its ability to account for uncertainty in other model parameters has not been exploited because of the heavy computational costs. With the increased computer power now available, the inverse problem and parameter estimation field should give serious consideration to applying it on a regular basis.

Acknowledgement

The authors wish to acknowledge the financial support of Sandia National Laboratories and the technical support of Dr. K. J. Dowding.

REFERENCES

1. E. L. Lehmann and G. Casella, *Theory of Point Estimation,* Springer, New York, 1998.
2. V. F. Turchin and V. Z. Nozik, Statistical regularization of the solution of incorrectly posed problems. *Izv. Atmospheric Oceanic Phys.* (1969) **5**, 29-38.
3. B. D'Ambrogi, S. Mäenpää and M. Markkanen, Discretization independent retrieval of atmospheric ozone profile. *Geophysica* (1999). **35** (1-2), 87-99.
4. F. P. Incropera and D. P. DeWitt, *Introduction to Heat Transfer*, J. Wiley and Sons, New York, 2001.
5. G. W. Burns and M. G. Scroger, The Calibration of Thermocouples And Thermocouple Materials, *NIST, SP-250-35*, Washington DC, 1989.

* Since k, eqn.(3), is effectively the ratio of two random quantities, even if both are normal, $p(\hat{k})$ is not normal, see Figure 3, and is not particularly easy to evaluate.

6. C-L Cheng and J. W. Van Ness, *Statistical Regression with Measurement Error,* Oxford Univ. Press, Oxford, UK, 1999.

7. A. Zellner, *An Introduction to Bayesian Inference in Econometrics,* J. Wiley and Sons, New York, 1996.

8. G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, J. Wiley and Sons, New York, 1989.

9. A. Papoulis and S. V. Pillai, *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill, New York, 2002.

10. R. Shiavi, *Introduction to Applied Statistical Signal Analysis,* Academic Press, New York, 1999.

11. A. F. Emery, B. F. Blackwell and K. J. Dowding, The relationship between information, sampling rates, and parameter estimation models. *ASME J. Heat Transfer.* (2002) **124** (6), 1192-1199.

12. Y. Pawatin, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford Univ. Press, Oxford, UK, 2001.

13. H. W. Engl, M. Hanke and A. Neubauer, *Regularization of Inverse Problems,* Kluwer Academic Publ., Boston, MA, 1996.

14. B. Hofmann, *Regularization for Applied Inverse and Ill-posed Problems : A Numerical Approach,* B.G. Teubner, Leipzig, 1986.

15. M. H. J. Gruber, *Regression Estimators : A Comparative Study,* Academic Press, Boston, MA, 1990.

16. A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin, *Bayesian Data Analysis,* Chapman & Hall/CRC Press, Boca Raton, Fl, 2004.

17. J. Wang and N. Zabaras, Using Bayesian Statistics in the estimation of heat source in radiation. *Int. J. Heat Mass Transfer* (2005) **48**, 15-29.

18. A. O'Hagan and J. Forster, *Bayesian Inference, Kendall's Advanced Theory of Statistics,* Vol. **2b**, Oxford Univ. Press, Oxford, UK, 2004.

19. W. M. Bolstad, *Introduction To Bayesian Statistics,* Wiley-Interscience, New York, 2004.

20. A. Wald, *Statistical Decision Functions,* J. Wiley and Sons, New York, 1950.

21. A. Pazman, *Foundations of Optimal Design,* Reidel, Boston, MA, 1986.

22. J. Sacks, S. B. Schiller and W. J. Welch, Designs for computer experiments. *Technometrics* (1989) **31** (1), 41-47.