

DATA AND MODEL ASSIMILATION USING A BAYESIAN METHODOLOGY: MARKOV CHAIN MONTE CARLO SIMULATION

A. Roy and P. G. Georgopoulos

3rd CRESP Annual Meeting

June, 1998

CRESP/EOHSI Exposure Assessment Task Group

Environmental and Occupational Health Sciences Institute
a joint project of UMDNJ – R. W. Johnson Medical School
and Rutgers University
170 Frelinghuysen Road, Piscataway
New Jersey 08854

ABSTRACT

Bayesian methodologies offer a convenient means of combining two disparate sources of information, models and data, both of which are integral to the risk assessment process. Mechanistic environmental and biological process models are being increasingly applied in the assessment of risk because of their many advantages-notable among which is the generally greater confidence, relative to empirical models, of predictions beyond the range of available data. Another advantage of mechanistic models is that some information regarding parameter values is usually available, since these parameters are generally physically meaningful quantities. However, since the number of these parameters is typically large, the estimation of parameter values that are both physically reasonable and consistent with the data is a non-trivial exercise. The efficacy of conventional parameter estimation techniques drops sharply as the number of parameters to be estimated increases, and only a limited amount of information regarding the probability distribution of the parameters is available. In contrast, procedures based on Bayesian methodologies can generate full probability distributions for for an arbitrary number of model parameters, which are both physically meaningful and are consistent with the data. Conceptually, probability distributions of model parameters may represent the lack of knowledge regarding the “true” value of a parameter, the natural variability of a parameter, or a combination of both. This presentation describes proposed work to implement and test specific Bayesian methods of assimilating information in models and data. Some examples employing a numerical Bayesian method, Markov Chain Monte Carlo (MCMC) Simulation, are presented.

INTRODUCTION (A)

Hypothesis

Bayesian methodologies can be used to combine prior information on an arbitrary number of parameters in complex biological and environmental process models, with the information content of data, to obtain probabilistic parameter estimates.

Motivation

The efficacy of conventional parameter estimation techniques drops sharply as the number of parameters to be estimated increases, particularly if the model is nonlinear in the parameters. This limitation forces drastic simplifications and assumptions to be made in estimating parameters in complex environmental and biological process models. Bayesian methods have the potential of circumventing many of these limitations.

INTRODUCTION (B)

Stakeholder Involvement

Regulatory agencies such as the EPA are promoting the use of mechanistic environmental and biological process models, and probabilistic risk assessment. Bayesian methodology offers advantages over conventional methods in both these areas. Limitations in the number of parameters that can be estimated by conventional methods represent a severe drawback for mechanistic models, since these models generally contain many more parameters than empirical models.

Novel Aspects of Project

Bayesian parameter estimation methods are able to accommodate the estimation of an arbitrary number of parameters, generally without a comparable decrease in efficacy. Furthermore, Bayesian methods are able to incorporate known information on the parameter in terms of a prior probability distribution. Conventional parameter estimation techniques can only accommodate upper and lower bound constraints on estimated parameters.

METHODS (A)

Bayesian Inference

- The fundamental difference between Bayesian and classical statistical inference is that in Bayesian inference parameters are regarded as random variables, whereas in classical inference, parameters are regarded as having a “true” but unknown value.
- Bayesian inference on model parameters (θ) given observed data (y) involves the estimation of either:
 - the posterior distribution of the parameters conditioned on the data, $P(\theta|y)$, or
 - the joint posterior distribution, $P(y, \theta) = P(\theta|y)P(y)$, which is proportional to $P(\theta|y)$, since the data y are given.
- The above distributions are determined using Bayes Theorem:

$$P(\theta|y) \propto P(y|\theta)P(\theta),$$

where $P(\theta)$ is the prior distribution, and $P(y|\theta)$ is the sampling distribution. The sampling distribution is known as the likelihood function when y is fixed, and θ is unknown.

METHODS (B)

Informative and Non-informative Prior Distributions

Specification of a prior distribution for each parameter to be estimated is required in Bayesian inference. This characteristic of Bayesian inference can be both an advantage and a disadvantage. It is an advantage when the prior distribution is known, since unlike in conventional parameter estimation, the shape of the prior probability distribution function is taken into account. In conventional methods, at best, only an acceptable range can be specified for the parameter. In such cases the prior distribution is termed *informative*.

However, even when no information is available with respect to the parameter being estimated, it is necessary to select a prior distribution for the parameters being estimated. In such cases it is desirable to select a prior distribution that will have the least influence on the posterior distribution, and will in effect let the data “speak for themselves.” Such prior distributions are termed *non-informative*, a common example of which is the uniform distribution which in effect provides upper and lower bounds for the parameter being estimated.

METHODS (C)

Hierarchical Models

A key advantage of Bayesian inference over conventional parameter estimation methods is the ease with which hierarchical models can be developed. Many environmental and biological process models are ideally represented as hierarchical models.

A typical example is the representation of inter- and intra-individual pharmacokinetic and pharmacodynamic variability, which conceptualizes the variance in pharmacokinetic and pharmacodynamic response observed for a population as having two components. The intra-individual component describes the variance in an individual's response about the individual's mean response, whereas the inter-individual component describes the variance among the mean individual mean responses in the population.

METHODS (D)

Markov Chain Monte Carlo Simulation

- Parameter estimation using Markov Chain Monte Carlo (MCMC) Simulation is a method of Bayesian inference, in which numerical representations of probability distribution functions (*pdf's*) are generated for model parameters. The mean, variance, and other moments for the parameters can be calculated from the numerically approximated *pdfs*.
- MCMC Simulation is the name given to a class of methods that can be used to generate numerical approximations of a joint posterior distribution that incorporates both data and *a priori* information on parameter values.
- Several algorithms have been proposed for the Markov Chain random walk process that is used to generate the joint posterior (or target) distribution, some of which are:
 - Metropolis algorithm
 - Metropolis-Hastings algorithm
 - Langevin algorithm
 - Gibbs Sampling algorithm
 - Metropolis-Hastings-Green algorithm

METHODS (E)

Gibbs Sampling

- The Gibbs sampler is a computationally efficient algorithm for generating a Markov Chain that converges to a joint posterior distribution, given conditional distributions of the constituent random variables.
- The first iteration of the Gibbs sampler produces a point in parameter space $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_d^{(1)})$ from an arbitrary initial point $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)})$, by generating the following d random variables:

$$\begin{aligned}\theta_1^{(1)} &\sim P(\theta_1 | \theta_2^{(1)}, \theta_3^{(0)}, \dots, \theta_d^{(0)}, \mathbf{y}) \\ \theta_2^{(1)} &\sim P(\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_d^{(0)}, \mathbf{y}) \\ &\vdots \\ \theta_d^{(1)} &\sim P(\theta_d | \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{d-1}^{(1)}, \mathbf{y})\end{aligned}$$

- This process is repeated to generate the sequence $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \dots, \boldsymbol{\theta}^{(N)}$, which tends to $P(\boldsymbol{\theta} | \mathbf{y})$ as $N \rightarrow \infty$.

METHODS (F)

Establishing Convergence of MCMC

- The convergence of the Markov Chain to a stationary distribution must be established before the generated distribution can be equated with the joint posterior distribution.
- A method of establishing convergence has been proposed by Gelman and Rubin (Stat. Sci. 7:457–511, 1992), which is analogous to ANOVA.
 - Multiple simulation sequences of n iterations each are performed, and variances of parameters of interest are calculated from realizations of the parameter within-sequences, and compared to between-sequence variances.
 - The within- and between-sequence variances should be equal, provided the distributions have converged.
 - Gelman and Rubin (1992) have proposed the following statistic to judge convergence:

$$\sqrt{\hat{R}} = \sqrt{\frac{(n-1)W + B}{nW}}$$

where W and B are the between- and within-sequence variances respectively. Convergence is indicated when $\sqrt{\hat{R}} < 1.2$.

RESULTS (A)

Population Toxicokinetics of Tetrachloroethylene*

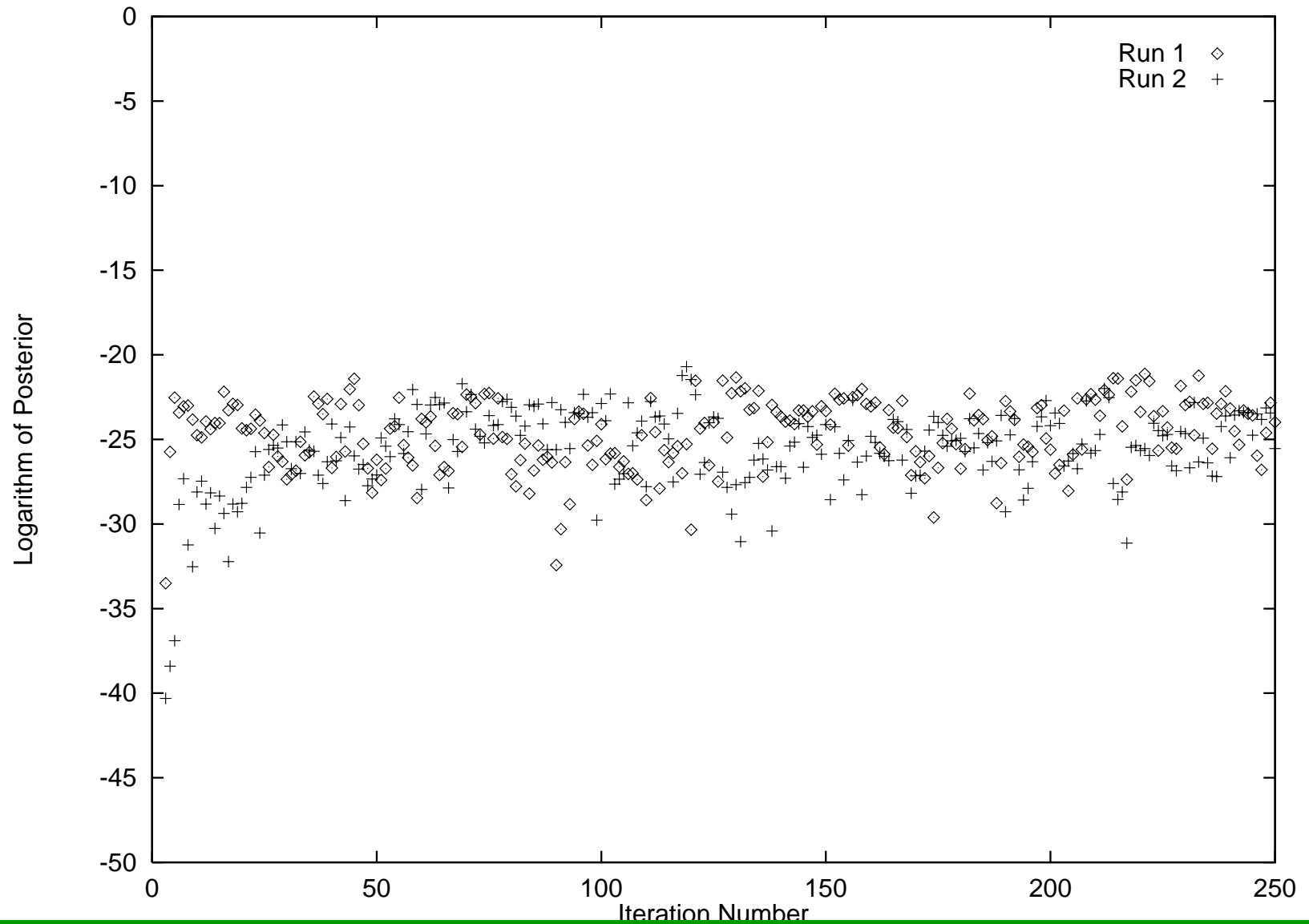
- The MCMC Simulation methodology has been applied to estimate the population parameters of a PBPK model for tetrachloroethylene (PERC) by Bois et al. (Arch. Tox., 1996), using human exposure data from Monster et al. (Int. Arch. Occup. Environ. Health, 1979).
- The joint posterior distribution that incorporates information from both the model and the data is: $P(\mathbf{y}, \boldsymbol{\theta}, \sigma^2 | \mathbf{E}, t, \mu, \Sigma^2)$ where,
 - \mathbf{w} vector of data
 - $\boldsymbol{\theta}$ vector of an individual's parameters
 - σ vector of within subject variability
 - \mathbf{E} vector of variables characterizing exposure level
 - t time
 - μ mean of population parameters
 - Σ variance of population parameters
- The above joint *pdf* can be expanded to:

$$P(\mathbf{y}, \boldsymbol{\theta}, \sigma^2 | \mathbf{E}, t, \mu, \Sigma^2) = P(\mathbf{y} | \boldsymbol{\theta}, \sigma^2, \mathbf{E}, t, \mu, \Sigma^2) P(\boldsymbol{\theta} | \sigma^2, \mathbf{E}, t, \mu, \Sigma^2) P(\sigma^2 | \mathbf{E}, t, \mu, \Sigma^2) P(\mu, \Sigma^2) .$$

*Summarized from: F. Bois, A. Gelman, J. Jiang, D. R. Maszle, L. Zeise, and G. Alexeef, Archives of Toxicology, 70, 347–355 (1996).

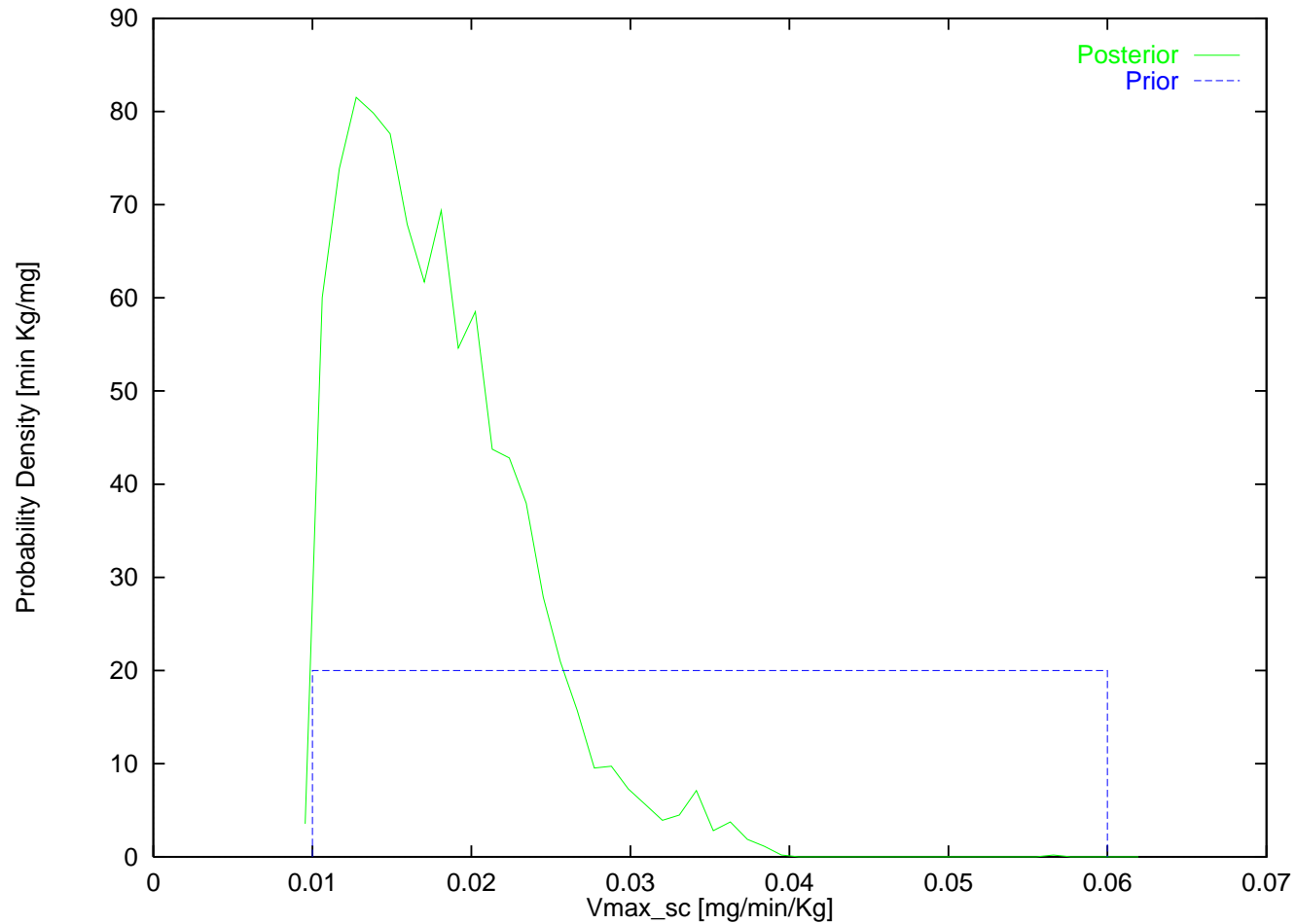
RESULTS (B)

Convergence of MCMC to a Stationary Distribution



RESULTS (C)

Comparison Between Prior and Posterior Distributions of Vmax



DISCUSSION

- The Markov Chain Monte Carlo (MCMC) simulation method with Gibbs sampling has been implemented using a physiologically based population pharmacokinetic (PB-PPK) model for perchloroethylene.
- MCMC simulations with the PERC PB-PPK model and data from Monster et al. (Int. Arch. Occup. Environ. Health, 1979) did converge to a stationary distribution, as judged by the criteria of Gelman and Rubin (Stat. Sci. 7:457–511, 1992).
- An approximately truncated lognormal distribution was predicted for V_{max} , the maximal rate of metabolism in the liver, based on a non-informative uniform prior distribution and the data of Monster et al. (Int. Arch. Occup. Environ. Health, 1979).

FUTURE PLANS

- The accuracy and precision of population pharmacokinetic data assimilation using Bayesian methodologies in general, and MCMC simulation with Gibbs sampling in particular, will be evaluated for assimilating data with a known PB-PPK models.
- Benzene or trichloroethylene (TCE) PB-PPK models will be used for the evaluation, since both of these chemicals are of concern at SRS. TCE due to an existing groundwater plume under the A- and M- Areas, and benzene due to intermittent atmospheric releases from high level waste vitrification processes will be considered.
- The potential application of the MCMC simulation method to other types of models, such as ecological foodweb models and groundwater transport models will also be investigated.