Lecture 9. The Use of Techniques within the Bayesian Framework of Statistics for the Solution of Inverse Problems

Helcio R. B. Orlande[1]
[1] Department of Mechanical Engineering, Politécnica/COPPE
Federal University of Rio de Janeiro, UFRJ, Cid. Universitária
Cx. Postal: 68503, Rio de Janeiro, RJ, 21941-972, Brazil
E-mail: helcio@mecanica.coppe.ufrj.br

Abstract. The solution of an inverse problem within the Bayesian framework is obtained by statistical inference on the posterior probability density. Such a density is obtained through Bayes' theorem and is proportional to the product of the likelihood function, which models the measurement errors, by the prior distribution, which models the information known about the parameters before the experimental data is available. The focus of this lecture is on Markov Chain Monte Carlo (MCMC) methods. Basic concepts, as well as practical issues regarding the implementation of MCMC methods, are presented. The Metropolis-Hastings algorithm, as well as its alternative version that samples the parameters by blocks, are described in detail. Monte Carlo methods usually involve large computational times. The Approximation Error Model technique and the Delayed Acceptance Metropolis-Hastings algorithm are thus presented, aiming at computational speed up and at making MCMC suitable for inverse problems of practical interest.

1. Introduction

The term *Bayesian* is commonly used to refer to techniques for the solution of inverse problems that fall within the framework of statistics developed by the Presbyterian minister Rev. Thomas Bayes (✫1702 - †1761) [1]. Such framework was actually established after Bayes' death, when his friend, Richard Price, published Bayes' famous paper, which dealt with the following problem: "*Given the number of times in which an unknown event has happened and failed: Required the chance that the probability of its happening in a single trial lies somewhere between two degrees of probability that can be named.*"[2]. On the other hand, it is attributed to Laplace the mathematical formulation that is known today as Bayes' theorem [3]. The term *Bayesian* was first used by R. A. Fisher, but in a pejorative context. Although born more than 120 years after the death of Bayes, Fisher was Bayes biggest intellectual rival [3]. The major issue by Fisher against Bayes and Laplace was that they used the concept of a *prior probability*, which represents the information about an unknown quantity before the measured data is available [3]. Fisher's theory relies solely on the measured data and on modelling of their associated uncertainty, aiming at unbiased inference and/or decision; therefore, it is usually referred to as the *frequentist* framework for statistics [1,3,4]. On the other hand, within the Bayesian framework, credit is also given to previous beliefs, in addition to that given to the measured data. Such previous information can even be qualitative but needs to be represented in terms of a probability distribution function, and regretfully induces bias in the results [1,3,4].

Nevertheless, the use of prior information in the Bayesian framework does not mean that it completely overtakes the information provided by the measured data, unless the last one is too uncertain to be really taken into account. Interestingly, one may also argue that life is Bayesian: think about life as a sequential process and notice that, at any day, our past beliefs are combined with new measured data, in order to provide at the end of the day a better understanding about different things of our interest, like physical/chemical phenomena, industrial processes, persons, or even the faster way to go to work.

Although not always considered in such a way, the solution of inverse problems can be appropriately formulated in terms of statistical inference [5]. Statistical inference refers to the process of drawing conclusions or making predictions based on limited information, beyond the immediate data that is available [4]. Note that this is exactly what is aimed with the solution of inverse problems, which can be broadly defined as those dealing with the estimation of unknown quantities appearing in the mathematical formulation of any kind of process, by using measurements of some dependent variable of the problem (observable response of the system) [5-27]. There are many techniques for the solution of inverse problems, but the most general ones are usually related to the minimization of an objective function that involves the difference between measured and estimated responses of the problem [5-27]. If the objective function is derived based on statistical hypotheses for the measurement errors and unknown parameters/functions, the minimization procedure can be related to statistical inference, thus resulting in point estimates for the unknowns that allow for estimations of their associated uncertainties [5,8]. Unfortunately, such is generally not the case, in special when the objective function is penalized with regularization terms.

The solution of inverse problems within the Bayesian framework is recast in the form of statistical inference from the so-called *posterior probability density*, which is the model for the conditional probability distribution of the unknown parameters given the measurements. The measurement model incorporating the related uncertainties is called the *likelihood*, that is, the conditional probability of the measurements given the unknown parameters. The model for the unknowns that reflects all the uncertainty of the parameters without the information conveyed by the measurements, is called the *prior* model [5,8,20,22,25-29]. The prior information can be combined with the likelihood to form the posterior distribution by using Bayes' theorem [5,8,20,22,25-29].

The objective of this text is to introduce some basic concepts regarding the solution of inverse heat transfer problems within the Bayesian framework. Special emphasis is given to the use of *Markov Chain Monte Carlo (MCMC) methods* [1,4,5,20,22,25-29]. Monte Carlo methods are also designated as *Stochastic Simulation techniques*, since values simulated (sampled) from the distribution of interest, which in general is not completely known, are used for the computation of its statistics [28]. Simulation techniques rely on probability results, such as the law of large numbers and the central limit theorem, which ensures that the approximate statistics tend to the actual ones as the number of simulated values increase [28].

This text is not aimed at a literature review about the subject, which would certainly include a very large number of works ranging from statistical, mathematical and computational aspects, to practical engineering applications. Indeed, an analysis of recent conferences on inverse problems clearly shows a trend of increasing number of papers that make use of solution techniques within the Bayesian framework, as faster computers become available. This text also does not cover Bayesian filters for the solution of state estimation problems.

The most complete source for the solution of inverse problems within the Bayesian framework is the book by Kaipio and Somersalo [5]. The reader is referred to the book by Gamerman and Lopes [28] and to the book edited by Brooks et al. [29] for deeper details about Markov Chain Monte Carlo methods. Fundamental material on Bayesian statistics can be found in the books by Lee [1] and Winkler [4]. A very didactical series of videos presenting the Monte Carlo Markov Chain methods can be found at https://www.youtube.com/watch?v=12eZWG0Z5gY. Two interesting books, with historical aspects and practical applications of Bayesian statistics in layman's terms, include references [3] and [30].

2. General Considerations

Consider the mathematical formulation of a heat transfer problem, which, for instance, can be linear or non-linear, one or multi-dimensional, involve a single or coupled heat transfer modes, etc. We denote the vector of parameters appearing in such formulation as

$$\mathbf{P}^T = [P_1, P_2, ..., P_N] \tag{1}$$

where N is the number of parameters. These parameters can possibly be thermal conductivity components, heat transfer coefficients, heat sources, boundary heat fluxes, etc. They can represent constant values of such quantities, or the parameters of the representation of a function in terms of known basis functions. For example, we can consider a heat source term $g_p(t)$ as a function of time, parameterized as follows:

$$g_p(t) = \sum_{j=1}^{N} P_j C_j(t) \tag{2.a}$$

where $C_j(t)$, $j = 1, \ldots, N$, are linearly independent basis functions that generate the space of the projected $g_p(t)$. Note that $C_j(t)$ can also be functions with local support, such as

$$C_j(t) = \begin{cases} 1 & , \quad for\ t_j < t < t_{j+1} \\ 0 & , \quad elsewhere \end{cases} \tag{2.b}$$

where the parameter P_j then represents the local value of the function in the time interval $t_j < t < t_{j+1}$, that is, $g_p(t_j) = P_j$, as illustrated by figure 1.

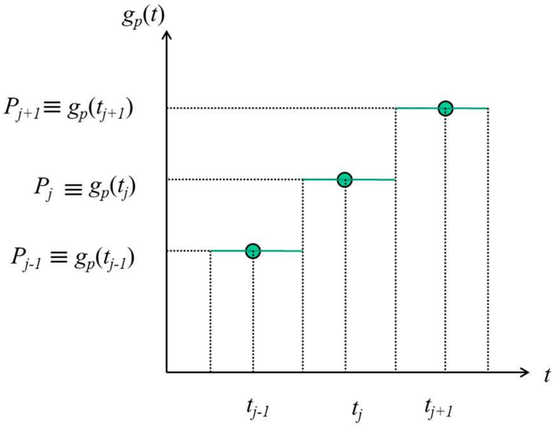


Figure 1. Parameters representing local values of a function that varies in time

Consider also that transient measurements are available within the medium, or at its surface, where the heat transfer processes are being mathematically formulated. The vector containing the measurements is written as:

$$\mathbf{Y}^T = \left(\vec{Y}_1 \ , \ \vec{Y}_2 \ , \ldots , \vec{Y}_I\right) \tag{3.a}$$

where $\vec{Y}_i$ contains the data of M sensors at time t_i, $i = 1, \ldots, I$, that is,

$$\vec{Y}_i = \left(Y_{i1} \ , \ Y_{i2} \ , \ldots , Y_{iM}\right) \quad \text{for } i=1, \ldots, I \tag{3.b}$$

Therefore, we have $D = MI$ measurements in total. Note that, in practice, the measured data are not limited to temperatures, but could also include heat fluxes, radiation intensities, etc.

Throughout this tutorial, the measurement errors are assumed to be additive, that is,

$$\mathbf{Y} = \mathbf{T}(\mathbf{P}) + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{T}(\mathbf{P})$ is the solution of the mathematical formulation of the physical problem, obtained with the vector of parameters $\mathbf{P}$, that is,

$$\mathbf{T}^T(\mathbf{P}) = [\vec{T}_1(\mathbf{P}) \ , \ \ \vec{T}_2(\mathbf{P}) \ , \cdots , \ \ \vec{T}_I(\mathbf{P})] \tag{5.a}$$

Where:

$$\vec{T}_i(\mathbf{P}) = [\,T_{i1}(\mathbf{P}) \ , \ \ T_{i2}(\mathbf{P}) \ , \cdots \ , \ \ T_{iM}(\mathbf{P})] \quad \text{for } i=1, \ldots, I \tag{5.b}$$

The mathematical formulation is supposed to perfectly represent the physical problem of interest. Similarly, the solution $\mathbf{T}(\mathbf{P})$ is supposed to be extremely accurate from the computational point of view. Anyhow, modelling errors can be appropriately taken into account within the Bayesian framework, as it will be apparent later in this text [5].

By further assuming that the measurement errors, $\boldsymbol{\varepsilon}$, are Gaussian random variables, with zero means, known covariance matrix $\mathbf{W}$ and independent of the parameters $\mathbf{P}$, their probability density function, $\pi(\boldsymbol{\varepsilon})$, is given by [5,8,20,22,25-29]:

$$\pi(\boldsymbol{\varepsilon}) = (2\pi)^{-D/2} \left|\mathbf{W}\right|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\varepsilon}^T \mathbf{W}^{-1}\boldsymbol{\varepsilon}\right\} \tag{6.a}$$

where $\pi = 4 \tan^{-1}(1)$. Due to the additive model for the measurement errors given by equation (4), equation (6.a) can be rewritten as

$$\pi(\boldsymbol{\varepsilon}) = \pi(\mathbf{Y}\left|\mathbf{P}\right.) = (2\pi)^{-D/2} \left|\mathbf{W}\right|^{-1/2} \exp\left\{-\frac{1}{2}[\mathbf{Y}-\mathbf{T}(\mathbf{P})]^T \mathbf{W}^{-1}[\mathbf{Y}-\mathbf{T}(\mathbf{P})]\right\} \tag{6.b}$$

which is the *likelihood function* for the above hypotheses regarding the measurement errors. The likelihood function gives the conditional probability density of different measurement outcomes $\mathbf{Y}$ with a fixed $\mathbf{P}$, which is denoted by $\pi(\mathbf{Y}\left|\mathbf{P}\right.)$ [5,8,20,22,25-29].

A very common approach for the solution of inverse problems, dealing with the estimation of the parameters $\mathbf{P}$ by using the measurements $\mathbf{Y}$, is to maximize the likelihood function. This can be accomplished through the minimization of the term inside the exponential function of equation (6.b), resulting in the following *maximum likelihood (ML) objective function*:

$$S_{ML}(\mathbf{P}) = \left[\mathbf{Y}-\mathbf{T}(\mathbf{P})\right]^T \mathbf{W}^{-1}\left[\mathbf{Y}-\mathbf{T}(\mathbf{P})\right] \tag{7}$$

The least squares norm can be obtained as a particular case of Eq. (7), if the measurements are uncorrelated and with constant variances σ^2 [8]. In this case, the covariance matrix $\mathbf{W}$ is given by:

$$\mathbf{W} = \sigma^2\mathbf{I} \tag{8}$$

where $\mathbf{I}$ is the identity matrix. Then, the minimization of Eq. (7) is equivalent to the minimization of the least squares norm:

$$S_{OLS}(\mathbf{P}) = \left[\mathbf{Y}-\mathbf{T}(\mathbf{P})\right]^T \left[\mathbf{Y}-\mathbf{T}(\mathbf{P})\right] \tag{9}$$

The covariance matrix of the values estimated for the parameters $\mathbf{P}$ with the minimization of equation (7), is given by [8]:

$$\mathrm{cov}(\mathbf{P}) = (\mathbf{J}^T\mathbf{W}^{-1}\mathbf{J})^{-1} \tag{10.a}$$

which reduces to

$$\mathrm{cov}(\mathbf{P}) = (\mathbf{J}^T\mathbf{J})^{-1}\sigma^2 \tag{10.b}$$

if $\mathbf{W}$ is given by equation (8). Equations (10.a, b) are exact for linear estimation problems, but can be used as approximations for nonlinear problems [8].

Therefore, in order to make use of the minimization of the least squares norm for obtaining point estimates for the parameters $\mathbf{P}$ that have some statistical meaning (for example, that allow estimates of the covariances of the estimated parameters with equation 10.b), all the statistical hypotheses stated above need to be valid [8]. Such a fact is quite often overlooked when an objective function is defined for the solution of an inverse problem via optimization techniques. Still, if the estimation problem is linear, the measurement errors are additive, with zero mean, and with a covariance matrix that is positive definite and known to within a multiplicative constant σ^2, that is,

$$\mathbf{W} = \hat{\mathbf{W}}\sigma^2 \tag{11}$$

the Gauss-Markov theorem [8,18] states that minimum variance estimates can be obtained with the minimization of

$$S_{GM}(\mathbf{P}) = \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right]^T \hat{\mathbf{W}}^{-1} \left[\mathbf{Y} - \mathbf{T}(\mathbf{P})\right] \tag{12}$$

even if the measurement errors are not Gaussian. In such a case, if $\hat{\mathbf{W}} = \mathbf{I}$, the minimization of the ordinary least squares norm provides minimum variance estimates. On the other hand, the covariance matrix of the values estimated for the parameters $\mathbf{P}$ cannot be computed with equations (10.a, b) since σ^2 is not known.

Different methods can be used for the minimization of equations (7), (9) or (12), after an analysis of the sensitivity coefficients of the parameters and an appropriate experimental design [8-26]. For a linear case, the minimization of equation (7) is obtained with:

$$\mathbf{P} = (\mathbf{J}^T\mathbf{W}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{W}^{-1}\mathbf{Y} \tag{13.a}$$

while, for the nonlinear case, the iterative procedure of Gauss' method gives:

$$\mathbf{P}^{k+1} = \mathbf{P}^k + (\mathbf{J}^T\mathbf{W}^{-1}\mathbf{J})^{-1}\mathbf{J}^T\mathbf{W}^{-1}[\mathbf{Y} - \mathbf{T}(\mathbf{P}^k)] \tag{13.b}$$

where the superscript k denotes the number of iterations and $\mathbf{J}$ is the sensitivity matrix.

We note that other maximum a posteriori objective functions can be derived if the measurement errors follow density functions different from the Gaussian distribution examined above.

3. Bayesian Framework

For the solution of inverse problems within the Bayesian framework, all variables included in the mathematical formulation of the physical problem are modelled as random variables.

Techniques for the solution of inverse problems within the Bayesian framework can be summarized in the following steps [5]:

1. Based on all information available for the parameters $\mathbf{P}$ before the measured data $\mathbf{Y}$ is taken, select a probability distribution function, $\pi(\mathbf{P})$, that appropriately represents the prior information.
2. Select the likelihood function, $\pi(\mathbf{Y}|\mathbf{P})$, that appropriately models the measurement errors and involves a relation between the observations and the mathematical model of the physical problem under picture (see, for example, equation 6.b).
3. Develop methods to explore the posterior density function, which is the conditional probability distribution of the unknown parameters given the measurements, $\pi(\mathbf{P}|\mathbf{Y})$.

The formal mechanism to combine the new information (measurements) with the previously available information (prior) is known as the Bayes' theorem [5,8,20,22,25-29]. Let $\mathbf{P}$ and $\mathbf{Y}$ be continuous random variables. Then, we can write [4]:

$$\pi(\mathbf{P}|\mathbf{Y}) = \frac{\pi(\mathbf{P},\mathbf{Y})}{\pi(\mathbf{Y})} \tag{14}$$

that is, the conditional density of the random variable $\mathbf{P}$ given a value of the random variable $\mathbf{Y}$ is the joint density of $\mathbf{P}$ and $\mathbf{Y}$ divided by the marginal density of $\mathbf{Y}$, where

$$\pi(\mathbf{Y}) = \int_{R^N} \pi(\mathbf{P},\mathbf{Y})\, d\mathbf{P} \tag{15}$$

The joint density $\pi(\mathbf{P},\mathbf{Y})$ is not generally known, but it can be written in terms of the likelihood and the prior as [4]:

$$\pi(\mathbf{P},\mathbf{Y}) = \pi(\mathbf{Y}|\mathbf{P})\pi(\mathbf{P}) \tag{16}$$

By substituting (16) into (14) we then obtain Bayes' theorem, which is given by:

$$\pi(\mathbf{P}|\mathbf{Y}) = \frac{\pi(\mathbf{Y}|\mathbf{P})\,\pi(\mathbf{P})}{\pi(\mathbf{Y})} \tag{17.a}$$

where $\pi_{posterior}(\mathbf{P}) = \pi(\mathbf{P}|\mathbf{Y})$ is the posterior probability density, $\pi(\mathbf{P})$ is the prior density, $\pi(\mathbf{Y}|\mathbf{P})$ is the likelihood function and $\pi(\mathbf{Y})$ is the marginal probability density of the measurements, which plays the role of a normalizing constant. Since the computation of $\pi(\mathbf{Y})$ with equation (15) is in general difficult, and usually not needed for practical calculations as will be apparent below, Bayes' theorem is commonly written as:

$$\pi_{posterior}(\mathbf{P}) = \pi(\mathbf{P}|\mathbf{Y}) \propto \pi(\mathbf{Y}|\mathbf{P})\pi(\mathbf{P}) \tag{17.b}$$

4. Maximum a Posteriori Objective Function

Consider a case with a Gaussian prior density model for the unknown parameters in the form:

$$\pi(\mathbf{P}) = (2\pi)^{-N/2} \left|\mathbf{V}\right|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{P}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{P}-\boldsymbol{\mu})\right] \tag{18}$$

where $\boldsymbol{\mu}$ and $\mathbf{V}$ are the known mean and covariance matrix for $\mathbf{P}$, respectively. By assuming normally distributed measurement errors, with zero means and known covariance matrix $\mathbf{W}$, additive and independent of the parameters $\mathbf{P}$, the likelihood function is given by equation (6.b). By substituting equations (6.b) and (18) into Bayes' theorem given by equation (17.b), we obtain:

$$\ln\left[\pi(\mathbf{P}\,|\,\mathbf{Y})\right] \propto -\frac{1}{2}\left[(D+N)\ln 2\pi + \ln|\mathbf{W}| + \ln|\mathbf{V}| + S_{MAP}(\mathbf{P})\right] \tag{19}$$

Where:

$$S_{MAP}(\mathbf{P}) = \left[\mathbf{Y}-\mathbf{T}(\mathbf{P})\right]^T \mathbf{W}^{-1}\left[\mathbf{Y}-\mathbf{T}(\mathbf{P})\right] + (\mathbf{P}-\boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{P}-\boldsymbol{\mu}) \tag{20}$$

Equation (19) reveals that the maximization of the posterior distribution can be obtained with the minimization of the objective function given by equation (20), denoted as the *maximum a posteriori* (MAP) *objective function* for the statistical hypotheses made above [5,8,20,22,25-29]. Equation (20) shows the contributions of the likelihood and of the prior distributions in this objective function, given by the first and second terms on the right-hand side, respectively. It is now interesting to notice that the maximum likelihood objective function (equation 7) is not a Bayesian estimator, since it does not contain information provided by the prior distribution for the parameters. Conspicuously, the least squares norm (equation 9) and other objective functions derived from equation (7), even those containing penalization terms (e.g., Tikhonov's regularization), are not Bayesian estimators, since they only explore the information provided by the measurements and, eventually, some characteristics of the parameters, like smoothness. Although the second term on the right-hand side of equation (20) is a quadratic form and resembles Tikhonov's regularization, there is a fundamental difference between the two approaches. Tikhonov's regularization focuses in obtaining a stabilized form of the original objective function and is not designed to yield error estimates that would have a statistical interpretation. In contrast, Bayesian inference assumes that the uncertainties in the likelihood and prior models reflect the actual uncertainties. Only if this condition is fulfilled, the uncertainties that are computed from equation (19) correspond to the actual posterior uncertainties [5].

Such as for the maximum likelihood objective function, different methods can be used for the minimization of equation (20) in order to obtain point estimates for the unknowns. For nonlinear problems, the Gauss method results in the following iterative procedure [5,8,20,22,25-29]:

$$\mathbf{P}^{k+1} = \mathbf{P}^k + [\mathbf{J}^T\mathbf{W}^{-1}\mathbf{J} + \mathbf{V}^{-1}]^{-1}\{\mathbf{J}^T\mathbf{W}^{-1}[\mathbf{Y}-\mathbf{T}(\mathbf{P}^k)] + \mathbf{V}^{-1}(\boldsymbol{\mu}-\mathbf{P}^k)\} \tag{21}$$

Note in equation (21) that with the MAP estimator, the conditioning of the matrix $\mathbf{J}^T\mathbf{W}^{-1}\mathbf{J}$ is improved with the matrix $\mathbf{V}^{-1}$, which is the inverse of the covariance matrix of the Gaussian prior information for the parameters. Therefore, the estimation of the parameters can be stabilized by using prior information with small covariances. Despite such desired effect for the regularization of the estimation procedure, the MAP estimator is biased and the expected value of $\mathbf{P}$ is $\boldsymbol{\mu}$ [8]. Such a fact clearly shows the important requirement of modeling the prior information as accurately as possible, for the success of the inverse analysis within the Bayesian framework. For a linear case, the covariance matrix of the posterior Gaussian distribution is given by [8]:

$$\operatorname{cov}(\mathbf{P}) = (\mathbf{J}^T\mathbf{W}^{-1}\mathbf{J} + \mathbf{V}^{-1})^{-1} \tag{22}$$

which can be used as an approximation for nonlinear cases.

5. Markov Chain Monte Carlo (MCMC) Methods

The Gaussian likelihood and the Gaussian prior examined in section 4 resulted in an expression for the posterior (equation 19) from which a MAP point estimate can be obtained for the parameters, provided that the minimum of equation (20) exists. In this particular case (Gaussian likelihood and Gaussian prior), the prior is *conjugate* to the likelihood [1,4,5,28]. A class Π of prior distributions is said to form a conjugate family if the posterior density is in the class Π for all $\mathbf{P}$, whenever the prior density is in Π [1]. Although this property is valid for many cases that involve continuous distributions, in special those that belong to the exponential family [1,28], the posterior probability distribution may not allow an analytical treatment if non-conjugate prior probability densities are assumed for the parameters. Moreover, whereas the computation of the MAP estimate is an optimization problem, that is,

$$\mathbf{P}_{MAP} = \arg\max_{\mathbf{P}\in R^N} \pi(\mathbf{P}\,|\,\mathbf{Y}) \tag{23}$$

other point and confidence estimate from the posterior distribution typically require numerical integration. For example, one common point estimate is the conditional mean defined as [5]:

$$\mathbf{P}_{CM} = E(\mathbf{P}) = \int_{R^N} \mathbf{P}\,\pi(\mathbf{P}\,|\,\mathbf{Y})\,d\mathbf{P} \tag{24}$$

where $E(.)$ denotes the expected value. In general, the dimension N of the parameter space is large enough to make the numerical integration in equation (24) impractical. Besides that, the computation of the normalizing constant in the denominator of $\pi(\mathbf{P}\,|\,\mathbf{Y})$ (see equations 14-17) already constitutes a challenging problem by itself.

For those cases that the posterior is not analytical and/or numerical integrations required for estimates are not practical, Markov Chain Monte Carlo (MCMC) methods can provide a solution of the inverse problem, so that inference on the posterior probability becomes inference on its samples [1,4,5,20,22,25-28]. For example, the Monte Carlo integration of equation (24) can be approximated by [5]:

$$\mathbf{P}_{CM} = E(\mathbf{P}) = \int_{R^N} \mathbf{P}\,\pi(\mathbf{P}\,|\,\mathbf{Y})\,d\mathbf{P} \approx \frac{1}{n}\sum_{t=1}^{n} \mathbf{P}^{(t)} \tag{25}$$

where $\mathbf{P}^{(t)}$, for $t = 1, \ldots, n$, are samples from $\pi(\mathbf{P}\,|\,\mathbf{Y})$. Markov Chain Monte Carlo methods are used to obtain such samples.

Due to the simplicity in the application of MCMC methods, such a technique for the solution of inverse problems has been recently becoming quite popular, being applied even for cases where a MAP estimate could be feasible. One clear disadvantage on the application of Monte Carlo methods is the large computational time required. On the other hand, the use of computationally fast reduced models for the physical problem can be appropriately accommodated within the Bayesian framework [5], so that the application of MCMC methods to many practical problems is nowadays possible.

Concepts and properties of Markov chains are presented in this section, which is then finished with a powerful, simple and popular MCMC algorithm. Some practical aspects and speedup techniques for the implementation of MCMC methods are delayed to other sections further below.

Markov Chains

The Markov chain is named after the Russian mathematician A. A. Markov, who developed such concept by investigating the alternance of vowels and consonants in a Russian poem. Poincaré also dealt with sequences of random variables that were in fact Markov chains [28]. A Markov chain is a stochastic process that, given the present state, past and future states are independent. The collection of the random quantities $\{\mathbf{P}^{(t)} : t \in T\}$ is said to be a stochastic process with state space S and index set T. The state space is a subset of R^N, that is, the support of the parameter vector, while here T is the set of Natural numbers that index the states of the Markov chain [28].

The stochastic process is a Markov chain if it satisfies the Markov condition [1,4,5,20,22,25-29]:

$$q(\mathbf{P}^{t+1} = \mathbf{y}\,|\,\mathbf{P}^{t} = \mathbf{x}, \mathbf{P}^{t-1} = \mathbf{x}^{t-1}, \ldots, \mathbf{P}^{0} = \mathbf{x}^{0}) = q(\mathbf{P}^{t+1} = \mathbf{y}\,|\,\mathbf{P}^{t} = \mathbf{x}) \text{ for all } \mathbf{y}, \mathbf{x}, \mathbf{x}^{t-1}, \ldots, \mathbf{x}^{0} \in S \tag{26}$$

where q is a transition probability. Some concepts regarding Markov chains are now presented. The reader shall consult references [1,4,5,20,22,25-29] for further details.

If the transition probability does not depend on t, that is, if

$$q(\mathbf{P}^{t+m+1} = \mathbf{y}\,|\,\mathbf{P}^{t+m} = \mathbf{x}) = q(\mathbf{P}^{t+1} = \mathbf{y}\,|\,\mathbf{P}^{t} = \mathbf{x}) \qquad \text{for all } m \in T \tag{27}$$

the Markov chain is said to be *homogenous* [22].

A distribution p^* is said to be a *stationary distribution* of a chain if, once the chain is in p^*, it stays in this distribution. Suppose now that $p^{(t)} \to p^*$ as $t \to \infty$ for any $p^{(0)}$, where $p^{(t)}$ is the distribution at state t of the chain. Then, p^* is the *equilibrium distribution* of the Markov chain and the chain is said to be *ergodic*.

Consider the sequence of states $\mathbf{x} \to \mathbf{k}_1 \to \mathbf{k}_2 \to \cdots \mathbf{k}_t \to \mathbf{y}$ so that the transition probabilities $q(\mathbf{k}_1 | \mathbf{x}) \neq 0,\ q(\mathbf{k}_2 | \mathbf{k}_1) \neq 0,\ \ldots,\ q(\mathbf{y} | \mathbf{k}_t) \neq 0$. Then, there is a sequence of states from $\mathbf{x}$ to $\mathbf{y}$ with a nonzero probability of occurring in the Markov chain. It is said that $\mathbf{x}$ and $\mathbf{y}$ communicate. If $\mathbf{y}$ and $\mathbf{x}$ also communicate through nonzero transition probabilities, it is said that these two states intercommunicate. If all states in S intercommunicate, then the state space is said to be *irreducible* under q. A Markov chain is *reversible* if $p(\mathbf{x})\, q(\mathbf{y} | \mathbf{x}) = p(\mathbf{y})\, q(\mathbf{x} | \mathbf{y})$.

The period of a state $\mathbf{x}$, denoted by d_x, is the largest common divisor of the set $\{m \geq 1 : p^{(m)}(\mathbf{x}, \mathbf{x}) > 0\}$. A state $\mathbf{x}$ is aperiodic if $d_x = 1$. A chain is *aperiodic* if all of its states are aperiodic.

Metropolis-Hastings Algorithm

The most common MCMC algorithms are the Gibbs Sampler and the Metropolis-Hastings algorithm [1,4,5,20,22,25-29]. The Gibbs Sampler is not presented here for the sake of brevity. The Metropolis-Hastings algorithm was first devised by Metropolis et al. [31] in 1953, who aimed at the calculation of the properties of substances composed of interacting molecules. It was, therefore, a work focused on statistical mechanics, not in statistics itself. Although the paper has five co-authors [31], only the name of the first author became popular to designate the developed algorithm, which was lately generalized by Hastings in 1970 [32]. In fact, there are some controversies about who actually contributed on the work by Metropolis et al. [33].

The reason for the introduction of the above concepts about Markov chains is for the statement of following result regarding the Metropolis-Hastings algorithm [22]: *Let p be a given probability distribution. The Markov chain simulated by the Metropolis-Hastings algorithm is reversible with respect to p. If it is also irreducible and aperiodic, then it defines an ergodic Markov chain with unique equilibrium distribution p.*

Unfortunately, it might not be possible to prove that the chain is irreducible and/or aperiodic for practical cases. In fact, parameters with linearly-dependent sensitivity coefficients generally result on periodic and correlated chains and an equilibrium distribution is not reached. Similarly to classical methods of parameter estimation, where the sensitivity coefficients directly influence the topology of the objective function based on the likelihood (see equation 7, for example) and a global minimum might not exist, such coefficients directly influence the posterior distribution, which is now sought via the implementation of a Markov chain. Therefore, the sensitivity coefficients need also to be carefully examined if the solution of the inverse parameter estimation problem is to be obtained within the Bayesian framework. In classical methods based on the maximum likelihood objective function, parameters with small and linearly dependent sensitivity coefficients are usually deterministically fixed, based on values known from previous experience and/or literature. In approaches within the Bayesian framework, uncertainties on such kind of parameters can be appropriately taken into account through their prior distribution functions. The analysis of the sensitivity coefficients reveals that parameters with small and/or linearly dependent sensitivity coefficients require informative prior distributions for the success of the estimation procedure.

The Metropolis-Hastings algorithm draws samples from a candidate density, such as in acceptance-rejection sampling [1]. The acceptance-rejection method is used to generate samples from a density $p(\mathbf{P}) = \tilde{p}(\mathbf{P}) / K$, where the normalizing constant K might be unknown, such as in the posterior distribution given by equations (17.a, b). Instead of sampling from $p(\mathbf{P})$, assume that there exists a candidate density $h(\mathbf{P})$ that is easy to simulate samples from, where $\tilde{p}(\mathbf{P}) \leq c\,h(\mathbf{P})$ and c is a constant. The following steps are then used to obtain a random variable $\hat{\mathbf{P}}$ from density $p(\mathbf{P})$ with the acceptance-rejection method [1]:

1. Generate a random variable $\mathbf{P}^*$ from the density $h(\mathbf{P})$;
2. Generate a random value $U \sim \mathrm{U}(0,1)$, which is uniformly distributed in (0,1);
3. If $U \leq \tilde{p}(\mathbf{P}) / c\,h(\mathbf{P})$, let $\hat{\mathbf{P}} = \mathbf{P}^*$. Otherwise, return to step 1.

The implementation of the Metropolis-Hastings algorithm starts with the selection of a candidate or proposal distribution $q(\mathbf{P}^* \mid \mathbf{P}^{(t)})$, which is used to draw a new candidate sample $\mathbf{P}^*$, given the current sample $\mathbf{P}^{(t)}$ of the Markov chain. Remind that, for the solution of the inverse problem within the Bayesian framework, one aims at simulating the posterior distribution $\pi_{posterior}(\mathbf{P}) \propto \pi(\mathbf{Y}|\mathbf{P})\pi(\mathbf{P})$ (see equation 17.b). Hence, the balance (reversibility) condition of the Markov chain of interest is given by:

$$\pi_{posterior}(\mathbf{P}^{(t)})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \pi_{posterior}(\mathbf{P}^*)\, q(\mathbf{P}^{(t)} \mid \mathbf{P}^*) \tag{28}$$

In order to avoid eventual cases that $\pi_{posterior}(\mathbf{P}^{(t)})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t)}) > \pi_{posterior}(\mathbf{P}^*)\, q(\mathbf{P}^{(t)} \mid \mathbf{P}^*)$, that is, the process moves from $\mathbf{P}^{(t)}$ to $\mathbf{P}^*$ more often than the reverse, a probability $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ is introduced in equation (28), so that [1]:

$$\pi_{posterior}(\mathbf{P}^{(t)})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t)})\, \alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \pi_{posterior}(\mathbf{P}^*)\, q(\mathbf{P}^{(t)} \mid \mathbf{P}^*) \tag{29}$$

Therefore,

$$\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi_{posterior}(\mathbf{P}^*)\, q(\mathbf{P}^{(t)} \mid \mathbf{P}^*)}{\pi_{posterior}(\mathbf{P}^{(t)})\, q(\mathbf{P}^* \mid \mathbf{P}^{(t)})}\right] \tag{30}$$

where $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)}) = 1$ when the balance condition is satisfied. Equation (30) is also called the Metropolis-Hastings ratio. Notice that, for the computation of equation (30), there is no need to know the normalizing constant that appears in the definition of the posterior distribution (see equations 17.a,b).

Equation (29) shows that the probability of moving from the sample at the current state, $\mathbf{P}^{(t)}$, to $\mathbf{P}^*$ is now given by $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$. In the Metropolis-Hastings algorithm, a candidate $\mathbf{P}^*$ is accepted, such as in the acceptance-rejection method described above, based on the probability $\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$. The Metropolis-Hastings algorithm can then be summarized in the following steps [1,4,5,20,22,25-29]:

1. Let $t = 1$ and start the Markov chain with sample $\mathbf{P}^{(1)}$ at the initial state.
2. Sample a candidate point $\mathbf{P}^*$ from a proposal distribution $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$.
3. Calculate the probability $\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$ with equation (30).
4. Generate a random value $U \sim \mathrm{U}(0,1)$, which is uniformly distributed in (0,1).
5. If $U \le \alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$, set $\mathbf{P}^{(t+1)} = \mathbf{P}^*$. Otherwise, set $\mathbf{P}^{(t+1)} = \mathbf{P}^{(t)}$.
6. Make $t = t + 1$ and return to step 2 in order to generate the sequence $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(n)}\}$.

In this way, a sequence is generated to represent the posterior distribution and inference on this distribution is obtained from inference on the samples $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(n)}\}$. We note that values of $\mathbf{P}^{(t)}$ must be ignored until the chain has not converged to equilibrium (the burn-in period).

The proposal distribution plays a fundamental role in the success of the Metropolis-Hastings algorithm. Typical choices for $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$ are presented below.

(i) Random Walk: In this case $\mathbf{P}^* = \mathbf{P}^{(t)} + \mathbf{\Psi}$, where $\mathbf{\Psi}$ is a vector of random variables with distribution $q_1(\boldsymbol{\psi})$. Therefore, $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = q_1(\mathbf{\Psi})$. If the proposal distribution is symmetric, that is, $q_1(\boldsymbol{\psi}) = q_1(-\boldsymbol{\psi})$ or $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = q(\mathbf{P}^{(t)} \,|\, \mathbf{P}^*)$, equation (30) reduces to

$$\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi_{posterior}(\mathbf{P}^*)}{\pi_{posterior}(\mathbf{P}^{(t)})}\right] \tag{31}$$

Thus, for this choice of the proposal density, equation (31) shows that in step 5 of the Metropolis-Hastings algorithm, the candidate point $\mathbf{P}^*$ is always accepted if the move leads to a region of higher posterior probability. Furthermore, the candidate point can also be accepted if $\pi_{posterior}(\mathbf{P}^*) < \pi_{posterior}(\mathbf{P}^{(t)})$ with probability $\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$, thus allowing that the state space be highly explored.

Uniform and Gaussian distributions are commonly used for $q_1(\boldsymbol{\psi})$. Consider one single component P_j of the vector $\mathbf{P}$. For the uniform random walk proposal one can write:

$$P_j^* = P_j^{(t)} + w_j(2u_j - 1) \tag{32.a}$$

where u_j is a random number with uniform distribution in (0,1), that is, $u_j \sim \mathrm{U}(0,1)$, while w_j is the maximum variation for the parameter at each state of the Markov chain P_j .

For the Gaussian random walk proposal, we have

$$P_j^* = P_j^{(t)} + r_j \tag{32.b}$$

where now r_j is a Gaussian random number with zero mean and standard deviation s_j.

<u>(ii) Independent Move:</u> This choice for this proposal density is of the kind $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = q_2(\mathbf{P}^*)$, that is, it does not depend on the current state $\mathbf{P}^{(t)}$. In this case, the proposal density $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$ can be conveniently selected as the prior density $\pi(\mathbf{P}^*)$. By utilizing equation (17.b), equation (30) is rewritten as

$$\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi(\mathbf{Y} \,|\, \mathbf{P}^*)\pi(\mathbf{P}^*)}{\pi(\mathbf{Y} \,|\, \mathbf{P}^{(t)})\pi(\mathbf{P}^{(t)})} \frac{\pi(\mathbf{P}^{(t)})}{\pi(\mathbf{P}^*)}\right] \tag{33.a}$$

Hence, the Metropolis-Hastings ratio is given by the ratio of the likelihoods, that is,

$$\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi(\mathbf{Y} \,|\, \mathbf{P}^*)}{\pi(\mathbf{Y} \,|\, \mathbf{P}^{(t)})}\right] \tag{33.b}$$

Similarly to the random walk proposal, candidates moving to regions of higher probability (in this case, the likelihood) are always accepted. Candidates in regions of lower likelihoods can be accepted with probability $\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$.

A Metropolis-Hastings algorithm with an adaptive proposal distribution was presented by Haario et al [34]. This algorithm is not Markovian, but results in ergodic distributions. In this adaptive algorithm, a Gaussian proposal with center at the sample of the current state, $\mathbf{P}^{(t)}$, is given by [34,35]:

$$q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \begin{cases} \mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{0.1^2}{N}\mathbf{I}\right) & t \le 2N \\ (1-\beta)\mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{2.38^2}{N}\mathbf{\Sigma}_t\right) + \beta\mathrm{N}\left(\mathbf{P}^{(t)}, \dfrac{0.1^2}{N}\mathbf{I}\right) & t > 2N \end{cases} \tag{34}$$

where $\mathrm{N}(\mathbf{a},\mathbf{B})$ is a Gaussian distribution with mean $\mathbf{a}$ and covariance matrix $\mathbf{B}$, N is the number of parameters, $\mathbf{I}$ is the identity matrix and $\mathbf{\Sigma}_t$ is the covariance matrix of the posterior distribution up to the state t. The positive constant β ($0 < \beta < 1$) is used to promote the mixing between $\mathrm{N}\left(\mathbf{P}^{(t)}, \frac{2.38^2}{N}\boldsymbol{\Sigma}_t\right)$ and $\mathrm{N}\left(\mathbf{P}^{(t)}, \frac{0.1^2}{N}\mathbf{I}\right)$, in order to avoid that the algorithm halt if $\boldsymbol{\Sigma}_t$ is not well defined.

Different modified versions of the Metropolis-Hastings algorithm can be found in the literature (see, for example, [29]). In particular, a modified version of the Metropolis-Hastings algorithm has been proposed for cases that involve groups of linearly dependent parameters [28,35]. In this modified version, the sampling procedure and the acceptance/rejection test are performed separately for each block of parameters, within one iteration of the Metropolis-Hastings algorithm [28,35]. As an example, consider a case where the vector of parameters $\mathbf{P}$ is split into two groups of parameters $\mathbf{P}_1$ and $\mathbf{P}_2$. The Metropolis-Hastings algorithm with sampling by block of parameters can then be summarized by the following steps:

1. Let $t=1$ and start the Markov chains with the sample $\mathbf{P}^{(1)}$.
2. Sample candidates $\mathbf{P}_1^*$ from the proposal distribution $q_1(\mathbf{P}_1^* \,|\, \mathbf{P}_1^{(t)})$ for the vector $\mathbf{P}_1$ and make $\mathbf{P}_2^* = \mathbf{P}_2^{(t)}$.
3. Compute the Metropolis-Hastings ratio
$$\alpha_1(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi(\mathbf{P}^* \,|\, \mathbf{Y})\, q_1(\mathbf{P}_1^{(t)} \,|\, \mathbf{P}_1^*)}{\pi(\mathbf{P}^{(t)} \,|\, \mathbf{Y})\, q_1(\mathbf{P}_1^* \,|\, \mathbf{P}_1^{(t)})}\right] \tag{35.a}$$
4. Generate a random number with a uniform distribution in (0,1), $U_1 \sim \mathrm{U}(0,1)$.
5. If $U_1 \le \alpha_1(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$, make $\mathbf{P}_1^{(t+1)} = \mathbf{P}_1^*$. Otherwise, make $\mathbf{P}_1^{(t+1)} = \mathbf{P}_1^{(t)}$.
6. Sample candidates $\mathbf{P}_2^*$ from the proposal distribution $q_2(\mathbf{P}_2^* \,|\, \mathbf{P}_2^{(t)})$ for the vector $\mathbf{P}_2$ and make $\mathbf{P}_1^* = \mathbf{P}_1^{(t+1)}$.
7. Compute the Metropolis-Hastings ratio
$$\alpha_2(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi(\mathbf{P}^* \,|\, \mathbf{Y})\, q_2(\mathbf{P}_2^{(t)} \,|\, \mathbf{P}_2^*)}{\pi(\mathbf{P}^{(t)} \,|\, \mathbf{Y})\, q_2(\mathbf{P}_2^* \,|\, \mathbf{P}_2^{(t)})}\right] \tag{35.b}$$
8. Generate a random number with a uniform distribution in (0,1), $U_2 \sim \mathrm{U}(0,1)$.
9. If $U_2 \le \alpha_2(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$, make $\mathbf{P}_2^{(t+1)} = \mathbf{P}_2^*$. Otherwise, make $\mathbf{P}_2^{(t+1)} = \mathbf{P}_2^{(t)}$.
10. Let $t=t+1$ and return to step 2 in order to generate the sequence $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, ..., \mathbf{P}^{(n)}\}$.

6. Practical Issues regarding Markov Chain Monte Carlo (MCMC) Methods

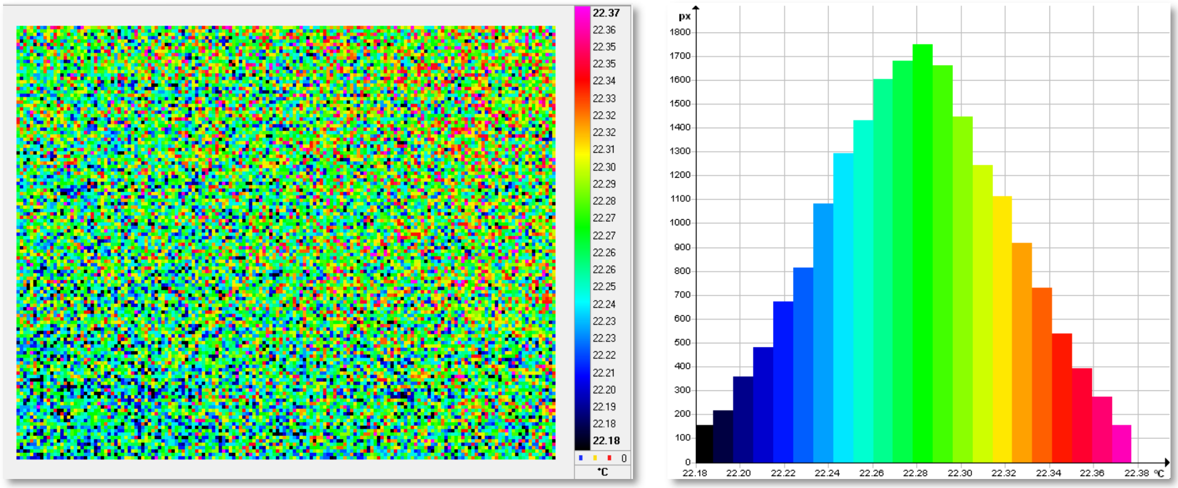
The objective of this section is to bring to the reader's attention some important aspects in the implementation of Markov Chain Monte Carlo methods. Although the discussion about likelihood and prior distributions is not limited to MCMC methods and is pertinent to Bayesian techniques in general, it was delayed until this section for the sake of organization of the text. Such is also the case regarding hierarchical models. In addition to these concepts, this section is also devoted to the analysis of the outputs of Markov chains.

Likelihood and Priors

The posterior distribution is proportional to the product of the likelihood function and the prior distribution (equation 17.b). As discussed in section 2, the likelihood function involves the solution of the mathematical formulation of the physical problem under analysis, that is, the solution of the direct or forward model, as well as the measurements and their related uncertainties. Measurement errors are modelled after the calibration of sensors and instruments used to collect the experimental data. The likelihood in section 2 was considered as Gaussian and given by equation (6.b). Such a model is in general appropriate for temperature measurements taken with thermocouples or infrared cameras. For example, figure 2.b presents the histogram of the readings (see figure 2.a) of a plate maintained at the constant temperature of 23 °C, obtained with a SC7600 Flir infrared camera [36]. This histogram clearly approximates a Gaussian distribution. For other likelihood models, appropriate to different physical phenomena, the reader is referred to [5].

A Gaussian prior was also considered in section 4, given by equation (18) for a multivariate case, with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$, denoted as $\mathbf{P} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{V})$. For one single parameter P_j , a *Gaussian prior* with mean μ_j and variance σ_j^2, $P_j \sim \mathrm{N}(\mu_j, \sigma_j^2)$**,** is given by

$$\pi(P_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2} \frac{(P_j - \mu_j)^2}{\sigma_j^2} \right] \qquad \text{in } -\infty < P_j < \infty \qquad (36)$$



(a) (b)

Figure 2. (a) Thermal image with an infrared camera of an isothermal plate;

(b) Histogram of the temperature measurements [36].

Random variables modelled by the Gaussian prior have support in R. Hence, they may assume negative values, although this might happen with small probabilities depending on the values of μ_j and σ_j^2. On the other hand, several physical parameters only allow positive values, such as, for example, thermal conductivity, specific heat and thermal diffusivity.

A very simple prior that allows lower and upper bounds for the parameter values is the *Uniform distribution* $P_j \sim \mathrm{U}(a,b)$ given by

$$\pi(P_j) = \begin{cases} \dfrac{1}{(b-a)} & , \quad a < P_j < b \\ 0 & , \quad elsewhere \end{cases} \tag{37}$$

Mean and variance for the uniform distribution are given by $\frac{1}{2}(a+b)$ and $\frac{1}{12}(b-a)^2$, respectively. In the uniform distribution, any value in $a < P_j < b$ is equally probable. If in this interval, values around a known mean are more likely to occur than elsewhere, like in a Gaussian distribution, but the probability density is zero in $P_j \le a$ and $P_j \ge b$, one possible prior can be obtained by combining equations (36) in (37), which is called *truncated Gaussian distribution*, that is,

$$\pi(P_j) = \begin{cases} \dfrac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\dfrac{1}{2} \dfrac{(P_j - \mu_j)^2}{\sigma_j^2} \right] & , \quad a < P_j < b \\ 0 & , \quad elsewhere \end{cases} \tag{38}$$

where $a < \mu_j < b$.

Other distributions that satisfy positive constraints are available. For example, the *Rayleigh distribution* $P_j \sim \mathrm{R}(\gamma_0)$ is given by

$$\pi(P_j) = \frac{P_j}{\gamma_0^2} \exp\left[-\frac{1}{2} \left(\frac{P_j}{\gamma_0} \right)^2 \right] \qquad \text{for } P_j > 0 \tag{39}$$

and depends only on the scale parameter (centerpoint) γ_0. The mean and the variance of Rayleigh's distribution are given by $\gamma_0 \sqrt{\frac{\pi}{2}}$ and $\frac{4-\pi}{2}\gamma_0^2$, respectively.

The *Gamma distribution* with parameters α and β, denoted as $P_j \sim \mathrm{G}(\alpha,\beta)$, has the following density:

$$\pi(P_j) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} P_j^{\alpha-1} \exp\left(-\frac{P_j}{\beta}\right) \quad \text{for} \quad P_j > 0 \tag{40}$$

with mean $\alpha\beta$ and variance $\alpha\beta^2$, where $\Gamma(\alpha)$ is the gamma function. For $\beta = 1$, the so-called one-parameter gamma distribution is obtained. The density that results by making $\alpha = 1$ is called exponential distribution.

The *Beta distribution* $P_j \sim \mathrm{Be}(\alpha,\beta)$ has support in $0 < P_j < 1$. The density of this distribution is given by

$$\pi(P_j) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} P_j^{\alpha-1}(1-P_j)^{\beta-1} \quad \text{in} \quad 0 < P_j < 1 \tag{41}$$

with mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Figure 3 illustrates the probability distributions $\mathrm{U}(0,1)$, $\mathrm{N}(0.5,0.5^2)$, $\mathrm{R}(0.5)$, $\mathrm{G}(1.5,1.5)$ and $\mathrm{Be}(1.5,1.5)$. These distributions were normalized by their maximum values to allow comparison among them.

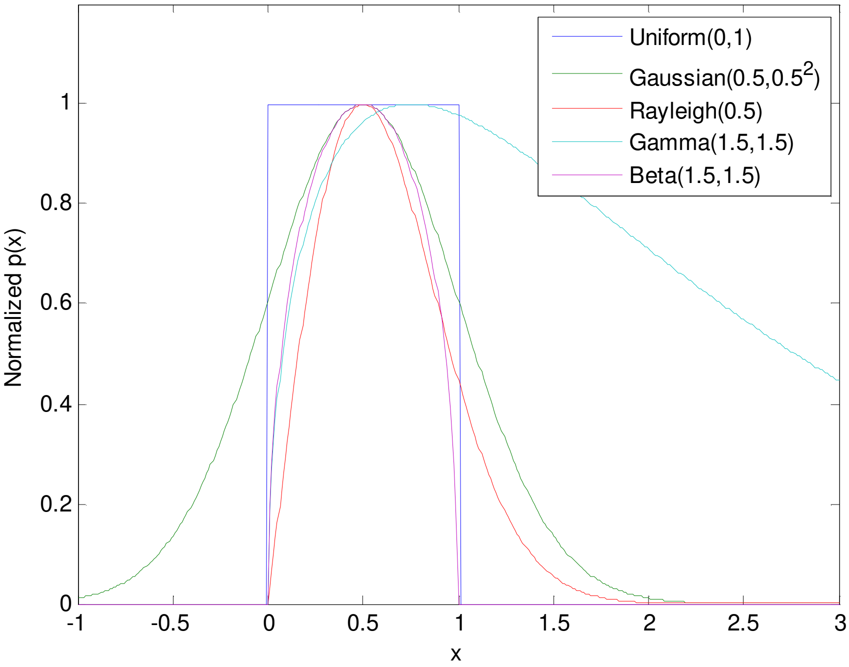


Figure 3. Probability distributions

The probability distributions given by equations (36) to (41) were written for one single random variable, but they can be easily extended for multivariate cases [1,4,5,8,28]. The multivariate Gaussian distribution is given by equation (18).

A multivariate prior is usually required for the solution of inverse problems in situations where the parameters represent point values of a function. Such is the case illustrated by figure 1 for time varying functions. Another typical case involves spatially distributed functions, like a thermophysical property that varies within the medium, where the parameter P_j is then associated to an average value of the function in a finite volume resulting from the discretization of the spatial domain. Markov Random Fields can be used to generate priors for these situations [5]. A collection $\{P_1, P_2, \ldots, P_N\}$ is a *Markov Random Field* if the full conditional distribution of P_j depends only on its set of neighbours [28].

A common use of a Markov Random Field is for priors that resemble Tikhonov's regularization [5], written in the following general form

$$\pi(\mathbf{P}) \propto \exp\left[-\frac{1}{2}\gamma\left\|\mathbf{D}(\mathbf{P}-\tilde{\mathbf{P}})\right\|^2\right] \tag{42}$$

where ||.|| denotes the L_2 norm. The constant γ is a parameter associated with uncertainties in the prior and $\tilde{\mathbf{P}}$ is a reference value for $\mathbf{P}$. The matrix $\mathbf{D}$ is such that each line of $\mathbf{D}(\mathbf{P}-\tilde{\mathbf{P}})$ involves the parameter P_j corresponding to that line and its neighbors, in order to characterize a Markov random field. For cases that $\mathbf{P}$ represent point values of a one-dimensional function (such as a function varying in time or in one single spatial coordinate), matrices like those used in Tikhonov's regularization serve well for this purpose. For example, one may use

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \quad \text{with size } (N-1)\,\mathrm{x}\,N \tag{43.a}$$

or

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{bmatrix} \quad \text{with size } (N-2)\,\mathrm{x}\,N \tag{43.b}$$

Equation (42) can be rewritten as

$$\pi(\mathbf{P}) \propto \exp\left[-\frac{1}{2}\gamma(\mathbf{P}-\tilde{\mathbf{P}})^T \mathbf{Z}(\mathbf{P}-\tilde{\mathbf{P}})\right] \tag{44.a}$$

where

$$\mathbf{Z} = \mathbf{D}^T\mathbf{D} \tag{44.b}$$

Equation (44.a) is in a form similar to that of a Gaussian distribution. For this reason, it is also called a Gaussian Markov Random Field [28] or a Gaussian Smoothness Prior [5]. By comparing equation (44.a) with the canonical Gaussian multivariate distribution, one can notice that the mean and the covariance matrix of this prior are given by $\tilde{\mathbf{P}}$ and $\gamma^{-1}\mathbf{Z}^{-1}$, respectively. Therefore, we can write the Gaussian Smoothness Prior as

$$\pi(\mathbf{P}) = (2\pi)^{-N/2}\gamma^{N/2}\left|\mathbf{Z}^{-1}\right|^{-1/2}\exp\left[-\frac{1}{2}\gamma(\mathbf{P}-\tilde{\mathbf{P}})^T \mathbf{Z}(\mathbf{P}-\tilde{\mathbf{P}})\right] \tag{45}$$

An important remark about this prior is that, with $\mathbf{D}$ given by equations (43.a,b), its variance is unbounded, since the matrix $\mathbf{Z}$ is singular and $\mathbf{Z}^{-1}$ does not exist. Densities with unbounded variances are denoted as *improper* [5,28].

We now discuss another Markov Random Field prior, which gives high probabilities for piecewise regular solutions with sparse gradients. The *Total Variation (TV) prior* satisfies these characteristics, being quite appropriate for spatially varying functions that contain large variations at few boundaries within the domain and with small variations within the regions limited by such boundaries [5]. The TV prior is given by [5]:

$$\pi(\mathbf{P}) \propto \exp\left[-\gamma TV(\mathbf{P})\right] \tag{46}$$

where

$$TV(\mathbf{P}) = \sum_{j=1}^{N} V_j(\mathbf{P}) \qquad V_j(\mathbf{P}) = \frac{1}{2}\sum_{i\in N_j} l_{ij}\left|P_i - P_j\right| \tag{47.a,b}$$

being N_j the set of neighbors to P_j and l_{ij} the length of the edge between neighbors.

The TV prior is improper, such as the Gaussian smoothness prior. The representation of equation (46) in terms of a canonical probability density would require the derivation of an expression for the normalizing constant $\int_{R^N}\pi(\mathbf{P})\,d\mathbf{P}$, or, at least, practical means for its computation. Although improper priors need to be used with caution, they do not pose difficulties for the application of the Metropolis-Hastings algorithm, since the normalizing constants of such densities are cancelled when $\alpha(\mathbf{P}^* \mid \mathbf{P}^{(t)})$ is computed with equation (30). On the other hand, both the Gaussian smoothness prior and the TV prior involve an additional parameter γ that needs to be specified for the application of MCMC methods. The specification of a value for such parameter can be made my numerical experiments, by using simulated experimental data that serve as a reference for the inverse problem under analysis.

On the other hand, within the Bayesian framework, if a parameter is not known it shall be regarded as part of the inference problem, leading to the use of hierarchical (*hyperprior*) models, as described below.

Hierarchical Models

The parameter γ appearing in the Gaussian smoothness prior given by equation (45) can be treated as a *hyperparameter*, that is, be estimated as part of the inference problem [5]. Consider, for example, the *hyperprior density* for γ in the form of a Rayleigh distribution (see equation 39), where the scale parameter γ_0 needs to be chosen in advance. Therefore, the posterior distribution, with the Gaussian likelihood given by equation (6.b), can be written as:

$$\pi(\gamma,\mathbf{P}|\mathbf{Y}) \propto \gamma^{(N+2)/2} \exp\left\{ -\frac{1}{2}[\mathbf{Y}-\mathbf{T}(\mathbf{P})]^T \mathbf{W}^{-1}[\mathbf{Y}-\mathbf{T}(\mathbf{P})] - \frac{1}{2}\gamma(\mathbf{P}-\tilde{\mathbf{P}})^T \mathbf{Z}(\mathbf{P}-\tilde{\mathbf{P}}) - \frac{1}{2}\left(\frac{\gamma}{\gamma_0}\right)^2 \right\} \tag{48}$$

On the other hand, the parameter γ appearing in the TV prior given by equation (46) cannot be treated as a hyperparameter. Such is the case because the normalizing constant of such prior is of difficult calculation and also depends on γ. Therefore, without the computation of the normalizing constant for this case, the effects of γ as a hyperparameter would not be correctly accounted for in the posterior distribution.

Output Analysis

We basically follow references [22,28] for the material presented in this section and consider the analysis on a single component P_j of the vector of parameters $\mathbf{P}$. Let $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ be a homogeneous and reversible Markov chain for P_j. A function $f(P_j^{(n)})$ from the sample $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ is called a *statistic* if it does not depend on any other unknown parameters. Some useful statistics are:

Minimum Value: $f(P_j^{(n)}) = P_{j,\min}^{(n)} = \min\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ (49.a)

Maximum Value: $f(P_j^{(n)}) = P_{j,\max}^{(n)} = \max\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ (49.b)

Median: $f(P_j^{(n)}) = \tilde{P}_j^{(n)} = \mathrm{med}\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ (49.c)

Mean: $f(P_j^{(n)}) = \bar{P}_j^{(n)} = \frac{1}{n}\sum_{t=1}^{n} P_j^{(t)}$ (49.d)

Variance: $f(P_j^{(n)}) = \mathrm{var}(P_j^{(n)}) = \frac{1}{n-1}\sum_{t=1}^{n}\left(P_j^{(t)} - \bar{P}_j^{(n)}\right)^2$ (49.e)

Since $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ are realizations of a random variable, a statistic is itself a random variable as well. A statistic of the sample will be a good representation of a statistic of the population if the sample is a good representation of the population. This certainly depends on the size n and on the independence of the individuals of the sample. Furthermore, since the sample $\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(n)}\}$ is obtained from a Markov chain, the chain should already have reached equilibrium before statistics can be computed to represent the solution of the inverse problem. For this reason, states of the Markov chain are discarded before the chain reaches equilibrium, which is called the burn-in period. If z states are needed for the chain to reach equilibrium, the sample used for the computation of the statistics is $\{P_j^{(z+1)}, P_j^{(z+2)}, \ldots, P_j^{(n)}\}$. The index of this sample is changed from $t = z+1, \ldots, n$ to $r = 1, \ldots, s$ for simplicity in the notation, where $s = n - z$ is the number of samples used for the computation of the statistics.

The *mean* of the sequence $P_j^{(r)} \equiv \left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$ is

$$\bar{P}_j^s = \frac{1}{s}\sum_{r=1}^{s} P_j^{(r)} \tag{50}$$

If the chain is ergodic, this mean provides a strongly consistent estimate of the mean of the limiting distribution, that is,

$$\bar{P}_j^s \to E\left[P_j\right] \quad \text{as} \quad s \to \infty \tag{51}$$

This result is the equivalent of the law of large numbers for a Markov chain.

If $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$ are independent samples, then the *variance of the mean* $\bar{P}_j^s$ is

$$\mathrm{var}[\bar{P}_j^s] = \frac{\mathrm{var}[P_j^{(r)}]}{s} \tag{52.a}$$

where $\mathrm{var}[P_j^{(r)}]$ is the variance of $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$. On the other hand, since the samples are in general correlated, equation (52.a) is rewritten as

$$\mathrm{var}[\bar{P}_j^s] = \frac{\tau_j \, \mathrm{var}[P_j^{(r)}]}{s} \tag{52.b}$$

where τ_j is the *integrated autocorrelation time* (IACT) for parameter P_j, which represents the number of correlated samples between independent samples in the chain $\left\{P_j^{(1)}, P_j^{(2)}, \ldots, P_j^{(s)}\right\}$. Therefore, the effective chain size, which gives the number of independent samples in the chain, is $s_{eff,j} = s / \tau_j$.

The *autocovariance function of lag k* of the chain for the parameter P_j is defined by:

$$C_j(k) = \mathrm{cov}[P_j^{(r)}, P_j^{(r+k)}] \tag{53}$$

Clearly, the variance of $P_j^{(r)}$ is $C_j(0)$.

The *normalized autocovariance function of lag k* is given by

$$\rho_j(k) = \frac{C_j(k)}{C_j(0)} \tag{54}$$

so that $\rho_j(0) = 1$, which means that $P_j^{(r)}$ is perfectly correlated with itself. The calculation of the normalized autocovariance function is straightforward, since several computational packages have functions available for such a purpose.

The integrated autocorrelation time is related to the *normalized autocovariance function* by

$$\tau_j = 1 + 2\sum_{k=1}^{\infty} \rho_j(k) \tag{55}$$

For the calculation of τ_j, the summation in equation (55) needs to be truncated at a finite number of terms $s^* \le s$. In fact, $\rho_j(k)$ is expected to tend to zero as k increases, but it will be dominated by noise for large k. Therefore, s^* can be selected by increasing k until $\rho_j(k)$ approaches zero, thus avoiding the terms that are dominated by noise.

For s sufficiently large and for an uniformly ergodic chain, the distribution of $\dfrac{\overline{P}_j^s - E\left[P_j\right]}{\sqrt{\mathrm{var}[\overline{P}_j^s]}}$, where $\mathrm{var}[\overline{P}_j^s]$ is given by equation (52.b), tends to a standard Gaussian distribution, with zero mean and unitary standard deviation. Thus,

$$\frac{\overline{P}_j^s - E\left[P_j\right]}{\sqrt{\mathrm{var}[\overline{P}_j^s]}} \overset{d}{\rightarrow} \mathrm{N}(0,1) \quad \text{as} \quad s \rightarrow \infty \tag{56}$$

where $\overset{d}{\rightarrow}$ indicates that the distribution of the random variable on the left tends to the distribution on the right. Equation (56) is an statement of the central limit theorem of the distribution of $\overline{P}_j^s$. Therefore, the mean of the samples in the Markov chain can be calculated with related uncertainties as $\bar{P}_j^s \pm \eta\sqrt{\mathrm{var}[\bar{P}_j^s]}$, where η is a constant that defines the approximate confidence interval of $\bar{P}_j^s$ (η = 2.576 for a 99% confidence interval).

The statistical efficiency of the sampling algorithm can be assessed by examining τ_j for each parameter P_j , $j = 1,...,N$. Algorithms that result in small values of τ_j promote better sampling. For cases involving many parameters, the statistical efficiency can be examined with the integrated autocorrelation time of the posterior distribution $\pi(\mathbf{P}^{(r)}\,|\,\mathbf{Y})$, $r = 1,\dots,s$ [35].

Quantitative techniques are available for the analysis of the convergence of a Markov chain to an equilibrium distribution. Geweke´s technique [37] compares the means calculated with the samples from different ranges of states of the Markov chain. Let:

$$\bar{P}_j^a = \frac{1}{s_a}\sum_{r=1}^{s_a} P_j^{(r)} \quad \text{and} \quad \bar{P}_j^b = \frac{1}{s_b}\sum_{r=s^*}^{s} P_j^{(r)} \tag{57.a,b}$$

be the means calculated with s_a and s_b states, respectively. Geweke [37] recommends:

$$s^* = s - s_b + 1 \quad ; \quad s_a = 0.1s \quad ; \quad s_b = 0.5s \tag{57c-e}$$

For the convergence analysis, it is also recommended to repeat the sampling procedure by starting the Markov chains from different initial values. Gelman and Rubin [38] developed a method for inference on multiple chains, based on two steps: (i) An estimate is obtained for the posterior distribution with an initial Markov chain, which is then used to start new independent chains. The initial states for these new multiple chains must have a dispersion larger than that of the initial chain; (ii) The new multiple chains are then used for inference with analyses inter chains and within each chain. The posterior distribution simulated with the multiple chains exhibit a variability larger than that of the initial chain.

The multiple chains also allow a convergence analysis to an equilibrium distribution that represents the sought posterior. We consider the case of a parameter P_j , $j = 1, ..., N$. The variance of the means of m chains, each one with n states, is given by [38]:

$$\frac{B_j}{n} = \frac{1}{(m-1)}\sum_{k=1}^{m}\left(\bar{P}_j^k - \bar{P}_j\right)^2 \tag{58}$$

where $\bar{P}_j^k$ is the mean of the chain k, $k = 1,...,m$, and $\bar{P}_j$ is the mean of these means.

The mean of the m variances of the chains $k = 1,...,m$, is given by [38]:

$$W_j = \frac{1}{m(n-1)}\sum_{k=1}^{m}\sum_{s=1}^{n}\left(P_j^{(s),k} - \bar{P}_j^k\right)^2 \tag{59}$$

where $P_j^{(s),k}$ is the sample for P_j at the state s, $s = 1,\ldots,n$, of chain k, $k = 1,...,m$.

The variance of the posterior distribution simulated with the multiple chains for P_j is thus obtained as [38]:

$$\hat{\sigma}_j^2 = \left(1 - \frac{1}{n}\right) W_j + \frac{B_j}{n} \tag{60}$$

This variance of the mn samples of the multiple chains, $\hat{\sigma}_j^2$, overestimate the variance of the actual posterior, while the equilibrium distribution has not been reached. On the other hand, W_j underestimates the variance of the actual posterior if each chain has not reached equilibrium. Gelman and Rubin [38] thus proposed a parameter to indicate convergence based on $\hat{\sigma}_j^2$ and W_j, called scale reduction coefficient, which was simplified by Gamerman and Lopes [28], and is given by:

$$\hat{R}_j = \sqrt{\frac{\hat{\sigma}_j^2}{W_j}} \tag{61}$$

Note that $\hat{R}_j > 1$, but $\hat{R}_j \to 1$ when $n \to \infty$. Gelman and Shirley [39] have suggested the empirical test $\hat{R}_j < 1.1$ for convergence of the multiple chains, but larger threshold values have also been proposed [28].

7. Reduction of the Computational Time for Markov Chain Monte Carlo (MCMC) Methods

For many cases, the computation of the direct problem solution, needed for the solution of the inverse problem, is very time-consuming. Limitations are then imposed on the number of states of the Markov chain that can be computed within a feasible time, which can make the use of standard MCMC methods impractical, especially when the number of unknown parameters is large. One possible way to overcome such difficulties is to use model reduction or surrogate techniques, instead of the complete model, for the computation of the direct problem solution at each state of the Markov chain.

Since reduced or surrogate models do not exactly reproduce the associated complete formulation of direct problems, different approaches have been developed in order to improve the solution of inverse problems obtained with these approximate models. Among such approaches, we have the Delayed Acceptance Metropolis-Hastings (DAMH) algorithm [40] and the Approximation Error Model (AEM) [5,41-45]. In the DAMH algorithm [40], the Metropolis-Hastings (MH) algorithm is regularly applied with the reduced model. If a proposal sample is accepted with the reduced model, another test of Metropolis-Hastings is performed with the complete model, to finally decide if such sample should be accepted or not. In this sense, the DAMH can be seen as two nested Metropolis-Hastings algorithms, where the outer loop acts as a filter to pre-evaluate proposal candidates with the reduced model. In the AEM approach [5,41-45], the statistical model of the approximation error is constructed from the prior information and then represented as additional noise in the measurement model, for the solution of the inverse problem. It should be noted that there is a fundamental difference between the DAMH and the AEM approaches. While the AEM uses the posterior modified by the error of the reduced model, the DAMH generates samples from the correct posterior. Such two approaches were successfully applied to a three-dimensional inverse heat conduction problem in reference [46].

Delayed Acceptance Metropolis-Hastings (DAMH) Algorithm

The DAMH algorithm can be summarized as follows [40]:

1. Let $t=1$ and start the Markov chain with the sample $\mathbf{P}^{(1)}$ at the initial state.
2. Sample a candidate point $\mathbf{P}^*$ from a proposal distribution $q(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$.
3. Calculate the probability $\alpha_{red}(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$ by using the reduced model, where

$$\alpha_{red}(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi_{red}(\mathbf{P}^* \,|\, \mathbf{Y})\, q(\mathbf{P}^{(t-1)} \,|\, \mathbf{P}^*)}{\pi_{red}(\mathbf{P}^{(t-1)} \,|\, \mathbf{Y})\, q(\mathbf{P}^* \,|\, \mathbf{P}^{(t-1)})}\right] \tag{62.a}$$

4. Generate a random value $U_{red} \sim \mathrm{U}(0,1)$.
5. If $U_{red} \le \alpha_{red}(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$, proceed to step 6. Otherwise, return to step 2.
6. Calculate a new acceptance factor with the complete model

$$\alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)}) = \min\left[1, \frac{\pi(\mathbf{P}^* \,|\, \mathbf{Y})\, q(\mathbf{P}^{(t-1)} \,|\, \mathbf{P}^*)}{\pi(\mathbf{P}^{(t-1)} \,|\, \mathbf{Y})\, q(\mathbf{P}^* \,|\, \mathbf{P}^{(t-1)})}\right] \tag{62.b}$$

7. Generate a new random value $U \sim \mathrm{U}(0,1)$.
8. If $U \le \alpha(\mathbf{P}^* \,|\, \mathbf{P}^{(t)})$ set $\mathbf{P}^{(t+1)} = \mathbf{P}^*$. Otherwise, set $\mathbf{P}^{(t+1)} = \mathbf{P}^{(t)}$.
9. Make $t = t+1$ and return to step 2 in order to generate the sequence $\{\mathbf{P}^{(1)}, \mathbf{P}^{(2)}, \ldots, \mathbf{P}^{(n)}\}$.

where $\pi_{red}(\mathbf{P} \,|\, \mathbf{Y})$ and $\pi(\mathbf{P} \,|\, \mathbf{Y})$ are the posterior distributions with the likelihoods computed with the reduced model and with the complete model, respectively.

With the DAMH algorithm, it is expected to take advantage of the fast computations of the reduced model in order to find, in step 5, possible candidates to be accepted with the complete model in step 8. The DAMH algorithm can be quite effective, especially in the case of a low acceptance ratio of the Metropolis-Hastings algorithm. Therefore, depending on how fast the solution of the reduced model is as compared to that of the complete model, as well as on the acceptance ratio, the use of the DAMH algorithm might result in significant reductions in computational times, as compared to those from the regular Metropolis-Hastings algorithm applied with the complete model.

Approximation Error Model (AEM) Approach

In the approximation error model (AEM) approach, the statistical model of the approximation error is constructed and then represented as additional noise in the measurement model [5,41-45]. With the hypotheses that the measurement errors are additive and independent of the parameters $\mathbf{P}$, one can write

$$\mathbf{Y} = \mathbf{T}(\mathbf{P}) + \boldsymbol{\varepsilon} \tag{63}$$

where $\mathbf{T}(\mathbf{P})$ is the sufficiently accurate solution of the complete direct (forward) model. The vector of measurement errors, $\boldsymbol{\varepsilon}$, are assumed here to be Gaussian, with zero mean and known covariance matrix $\mathbf{W}$, so that the likelihood function is given by equation (6.b).

If the solution of a reduced model, $\mathbf{T}_{red}(\mathbf{P})$, is used for the solution of the inverse problem in place of the solution of the complete model, $\mathbf{T}(\mathbf{P})$, equation (63) becomes

$$\mathbf{Y} = \mathbf{T}_{red}(\mathbf{P}) + [\mathbf{T}(\mathbf{P}) - \mathbf{T}_{red}(\mathbf{P})] + \boldsymbol{\varepsilon} \tag{64}$$

By defining the error between the complete and the reduced model solutions as

$$\mathbf{e}(\mathbf{P}) = [\mathbf{T}(\mathbf{P}) - \mathbf{T}_{red}(\mathbf{P})] \tag{65}$$

equation (64) can be written as

$$\mathbf{Y} = \mathbf{T}_{red}(\mathbf{P}) + \boldsymbol{\eta}(\mathbf{P}) \tag{66}$$

where

$$\boldsymbol{\eta}(\mathbf{P}) = \mathbf{e}(\mathbf{P}) + \boldsymbol{\varepsilon} \tag{67}$$

One difficult with such an approach is to model the error $\boldsymbol{\eta}(\mathbf{P})$, which includes the direct problem solution errors, $\mathbf{e}(\mathbf{P})$, as well as the experimental errors, $\boldsymbol{\varepsilon}$. A simple, but very effective approximation error approach, is to model such an error as a Gaussian variable [5,41-45]. Another important point for the implementation of the approximation error model is that the statistics of $\boldsymbol{\eta}(\mathbf{P})$, like its mean and covariance matrix, are computed before the estimation procedure, based on the prior distribution [5,41-45]. Therefore, the use of the approximation error model with improper priors is not possible, since they exhibit unbounded variances. Consider, for instance, a Gaussian prior and a Gaussian likelihood, given by equations (18) and (6.b), respectively. By using the approximation error model approach, the posterior distribution is given by [41]:

$$\pi(\mathbf{P}|\mathbf{Y}) \propto \exp\left\{-\frac{1}{2}[\mathbf{Y}-\mathbf{T}_{red}(\mathbf{P})-\overline{\boldsymbol{\eta}}]^T \tilde{\mathbf{W}}^{-1}[\mathbf{Y}-\mathbf{T}_{red}(\mathbf{P})-\overline{\boldsymbol{\eta}}]-\frac{1}{2}(\mathbf{P}-\boldsymbol{\mu})^T \boldsymbol{\Gamma}^{-1}(\mathbf{P}-\boldsymbol{\mu})\right\} \quad (68)$$

where

$$\overline{\boldsymbol{\eta}} = \overline{\boldsymbol{\varepsilon}} + \overline{\mathbf{e}} + \boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}}\boldsymbol{\Gamma}^{-1}(\mathbf{P}-\boldsymbol{\mu}) \quad (69.a)$$

$$\tilde{\mathbf{W}} = \mathbf{W}_e + \mathbf{W} - \boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_{\mathbf{P}\boldsymbol{\eta}} \quad (69.b)$$

and $\overline{\boldsymbol{\varepsilon}}$ and $\overline{\mathbf{e}}$ are the means of $\boldsymbol{\varepsilon}$ and $\mathbf{e}$, respectively, while $\mathbf{W}_e$ is the covariance of $\mathbf{e}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}}$ is the covariance of $\boldsymbol{\eta}$ and $\mathbf{P}$. Equations (69.a,b) give the *complete error model* [41]. We note that, with the standard hypotheses regarding the measurement errors made above, $\overline{\boldsymbol{\varepsilon}} = 0$. By further neglecting the dependency of $\boldsymbol{\eta}$ and $\mathbf{P}$, that is, $\boldsymbol{\Gamma}_{\boldsymbol{\eta}\mathbf{P}} = 0$, equations (69.a,b) simplify to the so-called *enhanced error model*:

$$\overline{\boldsymbol{\eta}} \approx \overline{\mathbf{e}} \quad (70.a)$$

$$\tilde{\mathbf{W}} \approx \mathbf{W}_e + \mathbf{W} \quad (70.b)$$

References

[1] Lee, P. M., 2004, *Bayesian Statistics*, Oxford University Press, London.

[2] Bayes, T., 1763, An Essay towards Solving a Problem in the Doctrine of Chances, by the late Rv. Mr. Bayes, F.R.S. Communicated by Mr. Price in a Letter to John Cannon, A.M.R.F.S., *Phil. Trans.* 1763 53, 370-418, 1763, doi:10.1098/rstl.1763.0053

[3] Silver N., 2012, *The Signal and the Noise*, Penguin Press, New York.

[4] Winkler, R., 2003, *An Introduction to Bayesian Inference and Decision*, Probabilistic Publishing, Gainsville.

[5] Kaipio, J. and Somersalo, E., 2004, *Statistical and Computational Inverse Problems*, Applied Mathematical Sciences 160, Springer-Verlag.

[6] Beck, J. V., Blackwell, B. and St. Clair, C. R., 1985, *Inverse Heat Conduction: Ill-Posed Problems*, Wiley Interscience, New York.

[7] Tikhonov, A. N. and Arsenin, V. Y., 1977, *Solution of Ill-Posed Problems*, Winston & Sons, Washington, DC.

[8] Beck, J. V. and Arnold, K. J., 1977, *Parameter Estimation in Engineering and Science*, Wiley Interscience, New York.

[9] Alifanov, O. M., 1994, *Inverse Heat Transfer Problems*, Springer-Verlag, New York.

[10] Alifanov, O. M., Artyukhin, E. and Rumyantsev, A., 1995, *Extreme Methods for Solving Ill-Posed Problems with Applications to Inverse Heat Transfer Problems*, Begell House, New York.

[11] Woodbury, K., 2002, *Inverse Engineering Handbook*, CRC Press, Boca Raton.

[12] Sabatier, P. C., 1978, *Applied Inverse Problems*, Springer Verlag, Hamburg.

[13] Morozov, V. A., 1984, *Methods for Solving Incorrectly Posed Problems*, Springer Verlag, New York.

[14] Murio, D. A., 1993, *The Mollification Method and the Numerical Solution of Ill-Posed Problems*, Wiley Interscience, New York.

[15] Trujillo, D. M. and Busby, H. R., 1997, *Practical Inverse Analysis in Engineering*, CRC Press, Boca Raton.

[16] Hensel, E., 1991, *Inverse Theory and Applications for Engineers*, Prentice Hall, New Jersey.

[17] Kurpisz, K. and Nowak, A. J., 1995, *Inverse Thermal Problems*, WIT Press, Southampton, UK.

[18] Vogel, C., 2002, *Computational Methods for Inverse Problems*, SIAM, New York.

[19] Yagola A. G., Kochikov, I.V., Kuramshina, G. M. and Pentin, Y. A., 1999, *Inverse Problems of Vibrational Spectroscopy,* VSP, Netherlands.

[20] Calvetti, D., Somersalo, E., 2007, *Introduction to Bayesian Scientific Computing*, Springer, New York.

[21] Ozisik, M.N. and Orlande, H.R.B., 2000, *Inverse Heat Transfer: Fundamentals and Applications*, Taylor and Francis, New York.

[22] Tan, S., Fox, C., and Nicholls, G., 2006, *Inverse Problems, Course Notes for Physics 707*, University of Auckland.

[23] A. Tarantola, 1987, *Inverse Problem Theory*, Elsevier.

[24] M. Bertero, P. Boccacci, 1998, *Introduction to Inverse Problems in Imaging*, Institute of Physics.

[25] Orlande, H., Fudym, F., Maillet, D., Cotta, R., 2011, *Thermal Measurements and Inverse Techniques*, CRC Press, Boca Raton.

[26] Jari P. Kaipio & Colin Fox, 2011, The Bayesian Framework for Inverse Problems in Heat Transfer, *Heat Transfer Engineering*, 32:9, 718-753.

[27] Orlande, H. R. B., 2012, Inverse Problems in Heat Transfer: New Trends on Solution Methodologies and Applications, *Journal of Heat Transfer*, v.134, p.031011.

[28] Gamerman, D. and Lopes, H.F., 2006, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman & Hall/CRC, 2nd edition, Boca Raton.

[29] Brooks, S., Gelman, A., Jones, G., Meng, X, 2011, *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, Boca Raton.

[30] McGrayne, S. B., 2011, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, Devon.

[31] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953, Equation of State Calculation by Fast Computing Machines, *J. Chemical Phys.*, vol. 21, pp. 1087-1092

[32] Hastings, W. K., 1970, Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, vol. 57, pp. 97-109.

[33] http://en.wikipedia.org/wiki/Metropolis%E2%80%93Hastings_algorithm, consulted on September 06, 2019

[34] Haario, H., Saksman, E., Tamminen, J., 2001, An Adaptive Metropolis Algorithm, *Bernoulli*, vol. 7, pp. 223-242.

[35] Cui, T., 2010, *Bayesian Calibration of Geothermal Reservoir Models via Markov Chain Monte Carlo*, Ph.D. Thesis, The University of Auckland.

[36] Fonseca, H.M., Orlande, H.R.B., Fudym, O., Sepúlveda, F., 2014, A statistical inversion approach for local thermal diffusivity and heat flux simultaneous estimation, *Quantitative InfraRed Thermography*, pp. 170-189.

[37] Geweke, J., 1992, Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments, in *Bayesian Statistics*, Bernardo, J., Berger, J., Dawid, a., Smith, A., (eds)., Oxford University Press

[38] Gelman, A., Rubin, D., 1992, Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science*, vol. 7, pp. 457-472.

[39] Gelman, A., Shirley, K., 2011, *Inference from Simulations and Monitoring Convergence*, Chapter 6 in Brooks, S., Gelman, A., Jones, G., Meng, X., 2011, Handbook of Markov Chain Monte Carlo, CRC Press, Boca Raton

[40] Christen, J., Fox, C., Markov chain Monte Carlo Using an Approximation, Journal of Computational and Graphical Statistics, vol. 14, no. 4, pp. 795–810, 2005.

[41] Nissinen, A., 2011, *Modelling Errors in Electrical Impedance Tomography*, Dissertation in Forestry and Natural Sciences, University of Eastern Finland.

[42] Nissinen, A., Heikkinen, L., Kaipio. J., 2008, The Bayesian approximation error approach for electrical impedance tomography – experimental results, *Meas. Sci. Technology*, vol. 19., pp. 015501.

[43] Nissinen, A., Heikkinen, L.. Kolehmainen, V., Kaipio. J., 2009, Compensation of errors due to discretization, domain truncation and unknown contact impedances in electrical impedance tomography, *Meas. Sci. Technology*, vol. 20, pp. 105504.

[44] Nissinen, A., Kolehmainen, V., Kaipio. J., 2011, Compensation of modeling errors due to unknown boundary domain in electrical impedance tomography, *IEEE Trans. Med. Im*., vol. 30, pp. 231-242.

[45] Nissinen, A., Kolehmainen, V., Kaipio. J., 2011, Reconstruction of domain boundary and conductivity in electrical impedance tomography using the approximation error approach, *Int. J. Uncertainty Quant*., vol. 1, pp. 203–222.

[46] Orlande, H. R. B., Dulikravich, G. S., Neumayer, M., Watzenig, D., Colaço, M., 2014, Accelerated Bayesian Inference for the Estimation of Spatially Varying Heat Flux in a Heat Conduction Problem, *Numerical Heat Transfer, Part A: Applications*, vol. 65, pp. 1-25